Federated Minimax Optimization with Client Heterogeneity

Anonymous authors
Paper under double-blind review

Abstract

Minimax optimization has seen a surge in interest with the advent of modern applications such as GANs, and it is inherently more challenging than simple minimization. The difficulty is exacerbated by the training data residing at multiple edge devices or *clients*, especially when these clients can have heterogeneous datasets and heterogeneous local computation capabilities. We propose a general federated minimax optimization framework that subsumes such settings and several existing methods like Local SGDA. We show that naive aggregation of model updates made by clients running unequal number of local steps can result in optimizing a mismatched objective function – a phenomenon previously observed in standard federated minimization. To fix this problem, we propose normalizing the client updates by the number of local steps. We analyze the convergence of the proposed algorithm for classes of nonconvex-concave and nonconvex-nonconcave functions and characterize the impact of heterogeneous client data, partial client participation, and heterogeneous local computations. For all the function classes considered, we significantly improve the existing computation and communication complexity results. Experimental results support our theoretical claims.

1 Introduction

The massive surge in machine learning (ML) research in the past decade has brought forth new applications that cannot be modeled as simple minimization problems. Many of these problems, including generative adversarial networks (GANs) Goodfellow et al. (2014); Arjovsky et al. (2017); Sanjabi et al. (2018), adversarial neural network training Madry et al. (2018), robust optimization Namkoong & Duchi (2016); Mohajerin Esfahani & Kuhn (2018), and fair machine learning Madras et al. (2018); Mohri et al. (2019), have an underlying min-max structure. However, the underlying problem is often nonconvex, while classical minimax theory deals almost exclusively with convex-concave problems.

Another feature of modern ML applications is the inherently distributed nature of the training data Xing et al. (2016). The data collection is often outsourced to edge devices or *clients*. However, the clients may then be unable (due to resource constraints) or unwilling (due to privacy concerns) to share their data with a *central server*. Federated Learning (FL) Konečný et al. (2016); Kairouz et al. (2019) was proposed to alleviate this problem. In exchange for retaining control of their data, the clients shoulder some of the computational load, and run part of the training process locally, using only their own data. The communication with the server is infrequent, leading to further resource savings. Since its introduction, FL has been an active area of research, with some remarkable successes Li et al. (2020); Wang et al. (2021). Research has shown practical benefits of, and provided theoretical justifications for commonly used practical techniques, such as, multiple local updates at the clients Stich (2018); Khaled et al. (2020); Koloskova et al. (2020); Wang & Joshi (2021), partial client participation Yang et al. (2021), communication compression Hamer et al. (2020); Chen et al. (2021). Further, impact of heterogeneity in the clients' local data Zhao et al. (2018); Sattler et al. (2019), as well as their system capabilities Wang et al. (2020); Mitra et al. (2021) has been studied. However, all this research has been focused almost solely on simple minimization problems.

With its increasing usage in large-scale applications, FL systems must adapt to a wide range of clients. Data heterogeneity has received significant attention from the community. However, system-level heterogeneity remains relatively unexplored. The effect of client variability or *heterogeneity* can be controlled by forcing all the clients to carry out an equal number of local updates and utilize the same local optimizer Yu et al. (2019);

Table 1: Comparison of (**per client**) stochastic gradient complexity and the number of communication rounds needed to reach an ϵ-stationary solution (Definition 1), for different classes of nonconvex minimax problems. Here, n is the total number of clients. For a fair comparison with existing works, our results in this table are specialized to the case when all clients (i) have equal weights ($p_i = 1/n$), (ii) perform equal number of local updates ($\tau_i = \tau$), and (iii) use the same local update algorithm SGDA. See Table 2 for comparison under more general settings, when (i)-(iii) do not hold.

Work	Setting and Assumptions		Full Client Participation (FCP)	
	System Heterogeneity[a]	Partial Client Participation	Stochastic Gradient Complexity	Communication Rounds
Nonconvex-Strongly-concave (NC-SC)/Nonconvex-Polyak-Łojasiewicz (NC-PL): Theorem 1				
($n = 1$) Lin et al. (2020a)	-	-	$\mathcal{O}(1/\epsilon^4)$	-
Sharma et al. (2022)	✗	✗	$\mathcal{O}(1/(n\epsilon^4))$	$\mathcal{O}(1/\epsilon^3)$
Yang et al. (2022a)	✗	✓	$\mathcal{O}(1/(n\epsilon^4))$	$\mathcal{O}(1/\epsilon^2)$
Ours: (Corollary 1.2, Remark 3)	✓	✓	$\mathcal{O}\left(1/(n\epsilon^4)\right)$	$\mathcal{O}\left(1/\epsilon^2\right)$
Nonconvex-Concave (NC-C): Theorem 2				
($n = 1$) Lin et al. (2020a)	-	-	$\mathcal{O}(1/\epsilon^8)$	-
Sharma et al. (2022)	✗	✗	$\mathcal{O}(1/(n\epsilon^8))$	$\mathcal{O}(1/\epsilon^7)$
Ours: (Corollary 2.2)	✓	✓	$\mathcal{O}\left(1/(n\epsilon^8)\right)$	$\mathcal{O}\left(1/\epsilon^4\right)$
Nonconvex-One-point-concave (NC-1PC): Theorem 2				
Deng & Mahdavi (2021)	✗	✗	$\mathcal{O}(1/\epsilon^{12})$	$\mathcal{O}(n^{1/6}/\epsilon^8)$
Sharma et al. (2022)	✗	✗	$\mathcal{O}(1/\epsilon^8)$	$\mathcal{O}(1/\epsilon^7)$
Ours: (Remark 5)	✓	✓	$\mathcal{O}\left(1/(n\epsilon^8)\right)$	$\mathcal{O}\left(1/\epsilon^4\right)$

[a] Individual clients can run an unequal number of local iterations, using different local optimizers (see Section 4).

Haddadpour et al. (2019). However, this approach is inefficient if the client dataset sizes are widely different. Also, it would entail faster clients sitting idle for long durations Reisizadeh et al. (2022); Tziotis et al. (2022), waiting for stragglers to finish. Additionally, using the same optimizer might be inefficient or expensive for clients, depending on their system capabilities. Therefore, adapting to system-level heterogeneity forms a desideratum for real-world FL schemes.

Contributions. We consider a general federated minimax optimization framework, in the presence of both inter-client data and system heterogeneity. **System heterogeneity** means the participating clients can run an unequal number of local steps, and utilize different local solvers. We consider the problem

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

where f_i is the local loss of client i, p_i is the weight assigned to client i (e.g., the relative sample size), and n is the total number of clients. We study several classes of nonconvex minimax problems (see Table 1). Further,

- In our proposed algorithm, the participating clients may each perform different number of local steps, with different local optimizers. In this setting, naive aggregation of local model updates (as done in existing methods like Local Stochastic Gradient Descent Ascent) may lead to convergence in terms of a mismatched global objective (Corollaries 1.1, 2.1). We propose a simple normalization strategy to fix this problem.
- We achieve order-optimal or state-of-the-art computation complexity and significantly improve the communication complexity of existing methods (Corollaries 1.2, 2.2).
- Under the special case where all the clients (i) are assigned equal weights $p_i = 1/n$ in (1), (ii) carry out equal number of local updates ($\tau_i = \tau$ for all i), and (iii) utilize the same local-update algorithm, our results become directly comparable with existing work (see Table 1) and improve upon them as follows.
    1. For nonconvex-strongly-concave (NC-SC - Corollary 1.2) and nonconvex-PL (NC-PL - Remark 3) problems, our method has the order-optimal gradient complexity $\mathcal{O}(1/(n\epsilon^4))$. Further, we improve the communication from $\mathcal{O}(1/\epsilon^3)$ in Sharma et al. (2022) to $\mathcal{O}(1/\epsilon^2)$.[1]

[1] The recent work Yang et al. (2022a) proposes FSGDA algorithm and also achieves $\mathcal{O}(1/\epsilon^2)$ communication cost for NC-PL functions. However, our work is more general since we allow different number of local steps and different local solvers at the clients.

2. For nonconvex-concave (NC-C - Corollary 2.2) and nonconvex-one-point-concave (NC-1PC - Remark 5) problems, we achieve state-of-the-art gradient complexity, while significantly improving the communication costs from $\mathcal{O}(1/\epsilon^7)$ in Sharma et al. (2022) to $\mathcal{O}(1/\epsilon^4)$. For NC-1PC functions, we prove the linear speedup in gradient complexity with n that was conjectured in Sharma et al. (2022).
3. As an intermediate result in our proof, we prove the theoretical convergence of Local SGD for one-point-convex function minimization (see Lemma C.5 in Appendix C.4). The achieved convergence rate is the same as that shown for convex minimization in the existing literature Khaled et al. (2020).

It is worth pointing out that our proof technique is different from existing minimax literature (e.g., Sharma et al. (2022); Yang et al. (2022b)). With all the clients carrying out the same number of local steps, the existing federated analyses rely on virtual sequences of average iterates, to mimic the proof steps in centralized settings Lin et al. (2020a); Yang et al. (2022c). In our case, since different clients run different number of local steps, this strategy is no longer viable (see Remark 9).

2 Related Work

2.1 Single-client minimax

Nonconvex-Strongly-concave (NC-SC). To our knowledge, Lin et al. (2020a) is the first work to analyze a single-loop algorithm for stochastic (and deterministic) NC-SC problems. Although the $\mathcal{O}(\kappa^3/\epsilon^4)$ complexity shown is optimal in ϵ, the algorithm required $\mathcal{O}(\epsilon^{-2})$ batch-size. Qiu et al. (2020) utilized momentum to achieve $\mathcal{O}(\epsilon^{-4})$ convergence with $\mathcal{O}(1)$ batch-size. Recent works Yang et al. (2022c); Sharma et al. (2022) achieve the same rate without momentum. Yang et al. (2022c) also improved the dependence on the condition number κ. Second-order stationarity for NC-SC has been recently studied in Luo & Chen (2021). Lower bounds for this problem class have appeared in Luo et al. (2020); Li et al. (2021); Zhang et al. (2021).

Nonconvex-Concave (NC-C). Again, Lin et al. (2020a) was the first to analyze a single-loop algorithm for stochastic NC-C problems, proving $\mathcal{O}(\epsilon^{-8})$ complexity. In deterministic problems, this has been improved using nested Nouiehed et al. (2019); Thekumparampil et al. (2019) as well as single-loop Xu et al. (2020); Zhang et al. (2020) algorithms. For stochastic problems, Rafique et al. (2021) and the recent work Zhang et al. (2022) improved the complexity to $\mathcal{O}(\epsilon^{-6})$. However, both the algorithms have a nested structure, which at every step, solve a simpler problem iteratively. Achieving $\mathcal{O}(\epsilon^{-6})$ complexity with a single-loop algorithm has so far proved elusive.

2.2 Distributed/Federated Minimax

Recent years have also seen an increasing body of work in distributed minimax optimization. Some of this work is focused on decentralized settings, as in Rogozin et al. (2021); Beznosikov et al. (2021b,c); Metelev et al. (2022).

Of immediate relevance to us is the federated setting, where clients carry out multiple local updates between successive communication rounds. The relevant works which focused on convex-concave problems include Reisizadeh et al. (2020); Hou et al. (2021); Liao et al. (2021); Sun & Wei (2022). Special classes of nonconvex minimax problems in the federated setting have been studied in recent works, such as, nonconvex-linear Deng et al. (2020), nonconvex-PL Deng & Mahdavi (2021); Xie et al. (2021), and nonconvex-one-point-concave Deng & Mahdavi (2021). The complexity guarantees for several function classes considered in Deng & Mahdavi (2021) were further improved in Sharma et al. (2022). However, all these works consider specialized federated settings, either assuming full-client participation, or system-wise identical clients, each carrying out equal number of local updates. As we see in this paper, partial client participation is the most source of error in simple FL algorithms. Also, system-level heterogeneity can have crucial implications on the algorithm performance.

Comparison with Wang et al. (2020); Sharma et al. (2022); Yang et al. (2022a). Wang et al. (2022a) was, to our knowledge, the first work to consider the problem of system heterogeneity in simple

minimization problems, and proposed a normalized averaging scheme to avoid optimizing an inconsistent objective. Compared to Wang et al. (2020), we consider a more challenging problem and achieve higher communication savings (Table 1)[2]. Sharma et al. (2022) studied minimax problems in the federated setting but assumed an equal number of SGDA-like local updates, with full client participation. The recent work Yang et al. (2022a) considers NC-SC problem with full and partial client participation and achieves similar communication savings as ours. In comparison, our work considers a more general minimax FL framework with partial client participation, clients running an unequal number of local updates, and using different local solvers. Further, we analyze multiple classes of nonconvex-concave and nonconvex-nonconcave functions, improving the communication and computation complexity of existing minimax methods.

3 Preliminaries

Notations. We let $\|\cdot\|$ denote the Euclidean norm $\|\cdot\|_2$. Given a positive integer m, the set $\{1, 2, \ldots, m\}$ is denoted by $[m]$. Vectors at client i are denoted with subscript i, e.g., $\mathbf{x}_i$, while iteration indices are denoted using superscripts, e.g., $\mathbf{y}^{(t)}$ or $\mathbf{y}^{(t,k)}$. Given a function g, we define its gradient vector as $\left[\nabla_x g(\mathbf{x}, \mathbf{y})^\top, \nabla_y g(\mathbf{x}, \mathbf{y})^\top\right]^\top$, and its stochastic gradient as $\nabla g(\mathbf{x}, \mathbf{y}; \xi)$, where ξ denotes the randomness.

Convergence Metrics. In the presence of nonconvexity, we can only prove convergence to an *approximate* stationary point, which is defined next.

Definition 1 (ϵ-Stationarity)**.** A point $\mathbf{x}$ is an ϵ-stationary point of a differentiable function g if $\|\nabla g(\mathbf{x})\| \leq \epsilon$.

Definition 2. Stochastic Gradient (SG) complexity is the total number of gradients computed by all the clients during the course of the algorithm.

In special cases, where all the clients are weighted equally ($p_i = 1/n$, for all $i \in [n]$) and carry out equal number of local steps τ, we state the *per-client* gradient complexity for comparison with existing work. See Table 1 and Corollaries 1.2 and 2.2.

Definition 3 (Communication Rounds)**.** During a single communication round, the server sends its *global* model to a set of clients, which carry out multiple local updates starting from the same model, and return their *local* vectors to the server. The server then aggregates these local vectors to arrive at a new global model. Throughout this paper, we denote the number of communication rounds by T.

Next, we discuss some assumptions used in the paper.

Assumption 1 (Smoothness)**.** Each local function f_i is differentiable and has Lipschitz continuous gradients. That is, there exists a constant $L_f > 0$ such that at each client $i \in [n]$, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_1}$ and $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$,

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

Assumption 2 (Bounded Diameter)**.** The constraint set $\mathcal{Y}$ is convex and bounded.

Assumption 3 (*Local* Variance)**.** The stochastic gradient oracle at each client is *unbiased*. Also, there exist constants $\sigma_L, \beta_L \geq 0$ such that at each client $i \in [n]$, for all $\mathbf{x}, \mathbf{y}$,

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

Assumption 4 (*Global* Heterogeneity)**.** For any set of non-negative weights $\{w_i\}_{i=1}^n$ such that $\sum_{i=1}^n w_i = 1$, there exist constants $\beta_G \geq 1, \sigma_G \geq 0$ such that for all $\mathbf{x}, \mathbf{y}$,

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

[2]Under the conditions $p_i = 1/n, \tau_i = \tau$ for all i, for smooth minimization problems, Wang et al. (2020) requires $\mathcal{O}(1/\epsilon^3)$ communication rounds. For NC-SC problems (a harder problem class), we show an improved $\mathcal{O}(1/\epsilon^2)$ communication rounds.

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

If all f_i's are identical, we have $\beta_G = 1$, and $\sigma_G = 0$.

Most existing work uses simplified versions of Assumptions 3, 4, assuming $\beta_L = 0$ and/or $\beta_G = 0$.

4 Algorithm for Heterogeneous Federated Minimax Optimization

In this section, we propose a federated minimax algorithm to handle system heterogeneity across clients.

4.1 Limitations of Local SGDA

Following the success of FedAvg McMahan et al. (2017) in FL, Deng & Mahdavi (2021) was the first to explore a simple extension Local stochastic gradient descent-ascent (SGDA) in minimax problems. Between successive communication rounds, clients take multiple simultaneous descent/ascent steps to respectively update the min-variable $\mathbf{x}$ and max-variable $\mathbf{y}$. Subsequent work in Sharma et al. (2022) improved the convergence results and showed that LocalSGDA achieves optimal gradient complexity for several classes of nonconvex minimax problems. However, existing work on LocalSGDA also assumes the participation of all n clients in every communication round. More crucially, as observed with simple minimization problems Wang et al. (2020), if clients carry out an unequal number of local updates, or if their local optimizers are not all the same, LocalSGDA (like FedAvg) might converge to the stationary point of a different objective. This is further discussed in Sections 5.1 and 5.2, and illustrated in Figure 1, where the learning process gets disproportionately skewed towards the clients carrying out more local updates.

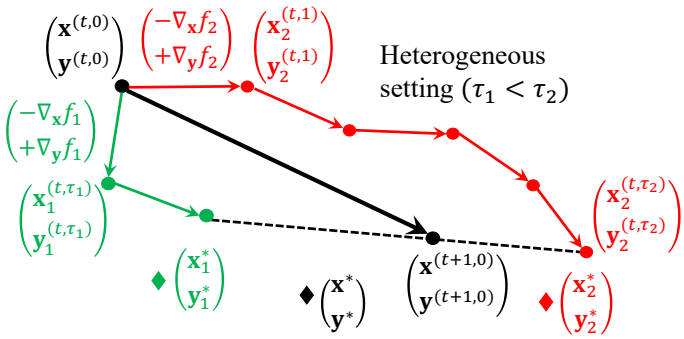

Figure 1: FedAvg with heterogeneous local updates. The green (red) triangle represents the local optimizer of $f_1(f_2)$, while $(\mathbf{x}^*, \mathbf{y}^*)$ is the global optimizer. The number of local updates at the clients is $\tau_1 = 2$, $\tau_2 = 5$.

Generalized Local SGDA Update Rule. To understand this mismatched convergence phenomenon with naive aggregation in local SGDA, recall that Local SGDA updates are of the form

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

where γ_x^s, γ_y^s are the server learning rates, $\Delta_{\mathbf{x},i}^{(t)} = \frac{1}{\eta_x^c}\left(\mathbf{x}_i^{(t,\tau_i^{(t)})} - \mathbf{x}^{(t)}\right)$, $\Delta_{\mathbf{y},i}^{(t)} = \frac{1}{\eta_y^c}\left(\mathbf{y}_i^{(t,\tau_i^{(t)})} - \mathbf{y}^{(t)}\right)$ are the scaled local updates. $\mathbf{x}_i^{(t,\tau_i^{(t)})}$ is the iterate at client i after taking $\tau_i^{(t)}$ local steps, and η_x^c, η_y^c are the client learning rates. Let us consider a generalized version of this update rule where $\Delta_{\mathbf{x},i}^{(t)}, \Delta_{\mathbf{y},i}^{(t)}$ are linear combinations of local stochastic gradients computed by client i, as $\Delta_{\mathbf{y},i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} a_i^{(t,k)} \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$, where $a_i^{(t,k)} \geq 0$. Commonly used client optimizers, such as, SGD, local momentum, variable local learning rates

can be accommodated in this general form (see Appendix A.1 for some examples). For this more general form, we can rewrite the $\mathbf{x}, \mathbf{y}$ updates at the server as follows

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

where $\mathbf{G}_{\mathbf{x},i}^{(t)} = [\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})]_{k=0}^{\tau_i^{(t)}} \in \mathbb{R}^{d_x \times \tau_i^{(t)}}$ contains the $\tau_i^{(t)}$ stochastic gradients stacked column-wise, $\bar{\boldsymbol{a}}_i^{(t)} = [a_i^{t,0}, a_i^{t,1}, \ldots, a_i^{t,\tau_i^{(t)}-1}]^\top$, $\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}$ are the normalized aggregates of the stochastic gradients and $\tau_{\text{eff}}^{(t)}$ is the *effective* number of local steps. Note that for simplicity, we assume that the constraint set $\mathcal{Y}$ has a large diameter. However, our algorithm can be easily modified to accommodate projection steps. Similar to

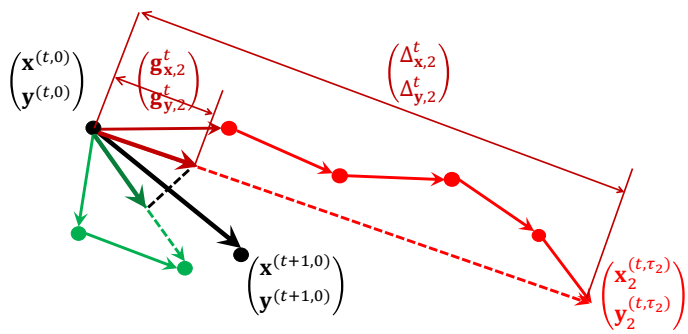

Figure 2: Generalized update rule in (2). Note that $(\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}) = \frac{1}{\tau_i}(\Delta_{\mathbf{x},i}^{(t)}, \Delta_{\mathbf{y},i}^{(t)})$. Also, at the server, the weighted sum $\sum_{i=1}^n w_i \mathbf{g}_{\mathbf{x},i}^{(t)}$ gets scaled by $\tau_{\text{eff}}^{(t)}$.

the observation for simple minimization problems in Wang et al. (2020), we see in Theorems 1, 2 that the resulting iterates of this general algorithm end up converging to the stationary point of a different objective $\widetilde{F} = \sum_{i=1}^n w_i f_i$. Further, in Corollary 1.1, we observe that this mismatch is a result of using weights w_i in (2) to weigh the clients' contribution.

4.2 Proposed Normalized Federated Minimax Algorithm

From the generalized update rule, we can see that setting the weights w_i equal to p_i will ensure that the surrogate objective $\tilde{F}$ matches with the original global objective F. Setting $w_i = p_i$ results in normalization of the local progress at each client before their aggregation at the server. As a result, we can preserve convergence to a stationary point of the original objective function F, even with heterogeneous $\{\tau_i^{(t)}\}$, as we see in Theorem 1 and Theorem 2.

The algorithm follows the steps given in Algorithm 1. In each communication round t, the server selects a client set $\mathcal{C}^{(t)}$ and communicates its model parameters $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ to these clients. The selected clients then run multiple local stochastic gradient steps. The number of local steps $\{\tau_i^{(t)}\}$ can vary across clients and across rounds. At the end of $\tau_i^{(t)}$ local steps, client i aggregates its local stochastic gradients into $\{\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}\}$, which are then sent to the server. Note that the gradients at client i, $\{\nabla f_i(\cdot, \cdot; \xi_i^{(t,k)})\}_{k=0}^{\tau_i^{(t)}}$, are normalized by $\|\bar{\boldsymbol{a}}_i^{(t)}\|_1$, where $\bar{\boldsymbol{a}}_i^{(t)} = [a_i^{t,0}, a_i^{t,1}, \ldots, a_i^{t,\tau_i^{(t)}-1}]^\top$ is the vector of weights assigned to individual stochastic gradients in

Algorithm 1 Fed-Norm-SGDA and Fed-Norm-SGDA+

Input: initialization $\mathbf{x}^{(0)}, \mathbf{y}^{(0)}$, Number of communication rounds T, learning rates: client $\{\eta_x^c, \eta_y^c\}$, server $\{\gamma_x^s, \gamma_y^s\}$, #local-updates $\{\tau_i^{(t)}\}_{i,t}$, S, $s = -1$
for $t = 0$ to $T-1$ **do**
  Server selects client set $\mathcal{C}^{(t)}$; sends them $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$
  if $t \bmod S = 0$ **then**
    $s \leftarrow s + 1$
    Server sends $\widehat{\mathbf{x}}^{(s)} = \mathbf{x}^{(t)}$ to clients in $\mathcal{C}^{(t)}$
  end if
  $\mathbf{x}_i^{(t,0)} = \mathbf{x}^{(t)}$, $\mathbf{y}_i^{(t,0)} = \mathbf{y}^{(t)}$ for $i \in \mathcal{C}^{(t)}$
  for $k = 0, \ldots, \tau_i^{(t)} - 1$ **do**
    $\mathbf{x}_i^{(t,k+1)} = \mathbf{x}_i^{(t,k)} - \eta_x^c a_i^{(t,k)} \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$
    $\mathbf{y}_i^{(t,k+1)} = \mathbf{y}_i^{(t,k)} + \eta_y^c a_i^{(t,k)} \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ # **y**-update for Fed-Norm-SGDA+
    $\mathbf{y}_i^{(t,k+1)} = \mathbf{y}_i^{(t,k)} + \eta_y^c a_i^{(t,k)} \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ # **y**-update for Fed-Norm-SGDA
  end for
  Client i aggregates its gradients to compute $\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}$
  $\mathbf{g}_{\mathbf{x},i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{a_i^{(t,k)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$
  $\mathbf{g}_{\mathbf{y},i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{a_i^{(t,k)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$
  $\mathbf{g}_{\mathbf{y},i}^{(t)} = \sum_{k=0}^{\tau_i^{(t)}-1} \frac{a_i^{(t,k)}}{\|\bar{\boldsymbol{a}}_i^{(t)}\|_1} \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$
  Clients $i \in \mathcal{C}^{(t)}$ communicate $\{\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}\}$ to the server
  Server computes aggregate vectors $\{\mathbf{g}_{\mathbf{x}}^{(t)}, \mathbf{g}_{\mathbf{y}}^{(t)}\}$ using (3)
  Server step: $\left\{\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \tau_{\text{eff}}^{(t)} \gamma_x^s \mathbf{g}_{\mathbf{x}}^{(t)}, \quad \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \tau_{\text{eff}}^{(t)} \gamma_y^s \mathbf{g}_{\mathbf{y}}^{(t)}\right.$
end for
Return: $\bar{\mathbf{x}}^{(T)}$ drawn uniformly at random from $\{\mathbf{x}^{(t)}\}_{t=1}^T$

the local updates.[3] The server aggregates these local vectors to compute global direction estimates $\mathbf{g}_{\mathbf{x}}^{(t)}, \mathbf{g}_{\mathbf{y}}^{(t)}$, which are then used to update the server model parameters $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$.

Client Selection. In each round t, the server samples $|\mathcal{C}^{(t)}|$ clients uniformly at random *without replacement* (WOR). While aggregating client updates at the server, client i update is weighed by $\tilde{w}_i = w_i n / |\mathcal{C}^{(t)}|$, i.e.,

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

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

5 Convergence Results

Next, we present the convergence results for different classes of nonconvex minimax problems. For simplicity, throughout this section we assume the parameters utilized in Algorithm 1 to be fixed across t. Therefore, $a_i^{(t,k)} \equiv a_i^{(k)}$, $\bar{\boldsymbol{a}}_i^{(t)} \equiv \boldsymbol{a}_i$, $\tau_i^{(t)} \equiv \tau_i$, $\tau_{\text{eff}}^{(t)} \equiv \tau_{\text{eff}}$ and $|\mathcal{C}^{(t)}| = P$, for all t.

[3]For LocalSGDA Deng & Mahdavi (2021); Sharma et al. (2022), $a_i^{(t,k)} = 1$ for all $i \in [n], t \in [T], k \in [\tau_i^{(t)}]$ and $\|\bar{\boldsymbol{a}}_i^{(t)}\|_1 = \tau_i^{(t)}$. Therefore, $\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}$ are simply the average of the stochastic gradients computed in the t-th round.

5.1 Non-convex-Strongly-Concave (NC-SC) Case

Assumption 5 (μ-Strong-concavity (SC) in $\mathbf{y}$)**.** A function f is μ-strong concave ($\mu > 0$) in $\mathbf{y}$ if

$$-f(\mathbf{x}, \tilde{\mathbf{y}}) \geq -f(\mathbf{x}, \bar{\mathbf{y}}) - \langle \nabla_y f(\mathbf{x}, \bar{\mathbf{y}}), \tilde{\mathbf{y}} - \bar{\mathbf{y}} \rangle + \frac{\mu}{2} \left\| \tilde{\mathbf{y}} - \bar{\mathbf{y}} \right\|^2, \qquad \text{for all } \mathbf{x} \in \mathbb{R}^{d_x}, \text{ and } \bar{\mathbf{y}}, \tilde{\mathbf{y}} \in \mathbb{R}^{d_y}.$$

General Convergence Result. We first show that the iterates of Algorithm 1 converge to the stationary point of a surrogate objective $\widetilde{F}$, where $\widetilde{F}(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^n w_i f_i(\mathbf{x}, \mathbf{y})$. $\{w_i\}_{i=1}^n$ are the aggregation weights used by the server (Line 19). See Appendix B for the full statement and proof.

Theorem 1. *Suppose the local loss functions $\{f_i\}_i$ satisfy Assumptions 1, 2, 3, 4, 5. Suppose the server selects $|\mathcal{C}^{(t)}| = P$ clients in each round t. Given appropriate choices of client and server learning rates, (η_x^c, η_y^c) and (γ_x^s, γ_y^s) respectively (see Appendix B.2), the iterates generated by Fed-Norm-SGDA satisfy*

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla \widetilde{\Phi}(\mathbf{x}^{(t)}) \right\|^2 \leq \underbrace{\mathcal{O}\left(\kappa^2 \sigma_G \sqrt{\frac{n-P}{n-1} \frac{E_w}{PT}} \right)}_{\substack{\textit{Partial participation} \\ \textit{error}}} + \underbrace{\mathcal{O}\left(\kappa^2 \sqrt{\frac{\Delta_{\widetilde{\Phi}} + A_w \sigma_L^2 + B_w \beta_L^2 \sigma_G^2}{P \tau_{\textit{eff}} T}} \right)}_{\textit{Error with full synchronization}} + \underbrace{\mathcal{O}\left(\kappa^2 \frac{C_w \sigma_L^2 + D \sigma_G^2}{\bar{\tau}^2 T} \right)}_{\textit{Local updates error}}. \tag{4}$$

where, $\kappa = L_f/\mu$ is the condition number, $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x}, \mathbf{y})$ is the envelope function, $\Delta_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}(\mathbf{x}^{(0)}) - \min_{\mathbf{x}} \widetilde{\Phi}(\mathbf{x})$, $\bar{\tau} = \frac{1}{n} \sum_{i=1}^n \tau_i$, $\tau_{\textit{eff}} = \sum_{i=1}^n p_i \|\boldsymbol{a}_i\|_1$, $A_w \triangleq n \tau_{\textit{eff}} \sum_{i=1}^n \frac{w_i^2 \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $B_w \triangleq n \tau_{\textit{eff}} \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $C_w \triangleq \sum_{i=1}^n w_i (\|\boldsymbol{a}_i\|_2^2 - [\alpha_i^{(t, \tau_i - 1)}]^2)$, $D \triangleq \max_i (\beta_L^2 \|\mathbf{a}_{i,-1}\|_2^2 + \|\mathbf{a}_{i,-1}\|_1^2)$, where $\mathbf{a}_{i,-1} \triangleq [a_i^{(0)}, a_i^{(1)}, \ldots, a_i^{(\tau_i - 2)}]^\top$ and $E_w \triangleq n \max_i w_i$.

Remark 1. The *first* term in (4) results from client subsampling ($P < n$). This explains its dependence on the data heterogeneity σ_G. The *second* term represents the optimization error for a centralized algorithm (see Appendix C.3 in Lin et al. (2020a)). The *last* term represents *client-drift*, the error if the client(s) run multiple local updates.

Theorem 1 states convergence for a surrogate objective $\widetilde{F}$. Next, we see convergence for the true objective F.

Corollary 1.1 (Convergence in terms of F)**.** *Given $\Phi(\mathbf{x}) \triangleq \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$, under the conditions of Theorem 1,*

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

where $\chi^2_{\mathbf{p} \| \mathbf{w}} \triangleq \sum_{i=1}^n \frac{(p_i - w_i)^2}{w_i}$, $\epsilon_{opt} \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla \widetilde{\Phi}(\mathbf{x}^{(t)}) \right\|^2$ denotes the optimization error in (4). If $p_i = w_i$ for all $i \in [n]$, then $\chi^2_{\mathbf{p} \| \mathbf{w}} = 0$. Also, then $\widetilde{F}(\mathbf{x}, \mathbf{y}) \equiv F(\mathbf{x}, \mathbf{y})$. Therefore, $\mathbf{y}^*(\mathbf{x}) = \arg\max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$ and $\widetilde{\mathbf{y}}^*(\mathbf{x}) = \arg\max_{\mathbf{y}} \widetilde{F}(\mathbf{x}, \mathbf{y})$ are identical, for all $\mathbf{x}$. Hence, (5) yields $\min_{t \in [T]} \left\| \nabla \Phi(\mathbf{x}^{(t)}) \right\|^2 \leq 2 \epsilon_{opt}$.

It follows from Corollary 1.1 that if in Algorithm 1, the server aggregation weights $\{w_i\}$ (Line 19) are the same as $\{p_i\}$, we get convergence to a stationary point of the true objective F. For the rest of this subsection, we assume $w_i = p_i$ for all $i \in [n]$.

In Table 2, we specialize the bound in (4) to SGDA-based local updates. We compare the bound under two cases: **Case 1**: equally-weighted clients ($p_i = 1/n$, for all i), all running $\tau_i \equiv \tau$ local updates; and **Case 2**: unequally weighted clients ($p_i \neq p_j$), running unequal local updates ($\tau_i \neq \tau_j$). The setting in **Case 1** has previously been considered in Sharma et al. (2022) (under full participation) and Yang et al. (2022a)[4]. Compared to Sharma et al. (2022), our bound has a smaller local-updates error term. This results in improved communication cost (see Corollary 1.2). The additional factors going from **Case 1** to the more general **Case 2** are highlighted in blue. The following insights can be drawn from Table 2.

[4]The condition number κ dependence is not explicitly stated in the results in Yang et al. (2022a).

Table 2: Comparison of convergence rates of Fed-Norm-SGDA (Theorem 1) and Fed-Norm-SGDA+ (Theorem 2), if all the clients run SGDA/SGDA+ based local updates, i.e., $a_i^{(t,k)} = 1$, for all i, k, t. The results are stated for (i) $(p_i = 1/n, \tau_i = \tau, \forall\, i \in [n])$; and (ii) $(p_i \neq p_j)$, $(\tau_i \neq \tau_j)$. The additional factors in (ii) relative to (i) are highlighted in blue. We state the results under partial-client participation (PCP). FCP results follow by choosing $P = n$. For simplicity, we assume uniformly bounded local variance ($\beta_L = 0$ in Assumption 3).

Nonconvex-Strongly-concave (NC-SC)/Nonconvex-Polyak-Łojasiewicz (NC-PL): (Theorem 1, Remark 3)

System Setting	Convergence Rate
$p_i = \frac{1}{n}, \forall\, i \in [n]$ $\tau_i = \tau, \forall\, i \in [n]$	Sharma et al. (2022) with $P = n$: $\mathcal{O}\left(\frac{\kappa^2\sigma_L^2}{\sqrt{n\tau T}} + \frac{\kappa^2 n\tau}{T}\left[\sigma_L^2 + \sigma_G^2\right]\right)$ Yang et al. (2022a) with $P < n$: $\mathcal{O}\left(\frac{\sigma_G^2}{\sqrt{PT}}\left(1 - \frac{P}{n}\right) + \frac{1}{\sqrt{PT}}\left[1 + \frac{\sigma_L^2}{\tau} + \sigma_G^2\right]\right)$ **Ours** with $P < n$: $\mathcal{O}\left(\kappa^2\sigma_G\sqrt{\frac{(n-P)}{(n-1)PT}} + \frac{\kappa^2\sigma_L}{\sqrt{P\tau T}} + \frac{\kappa}{T}\left[\frac{\sigma_L^2}{\tau} + \sigma_G^2\right]\right)$
Ours: $p_i \neq p_j,\ \tau_i \neq \tau_j$ $\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i$	$\mathcal{O}\left(\kappa^2\sigma_G\sqrt{\frac{(n-P)n\max_i p_i}{(n-1)PT}} + \frac{\kappa^2\sigma_L}{\sqrt{P\tau_{\text{eff}}T}}\sqrt{n\tau_{\text{eff}}\sum_{i=1}^n \frac{p_i^2}{\tau_i}} + \frac{\kappa}{T}\left[\frac{\sigma_L^2}{\bar{\tau}}\frac{\sum_{i=1}^n p_i\tau_i}{\bar{\tau}} + \sigma_G^2 \max_i \frac{\tau_i^2}{\bar{\tau}^2}\right]\right)$

Nonconvex-Concave (NC-C)/Nonconvex-One-Point-Concave (NC-1PC): (Theorem 2, Remark 5)

System Setting	Convergence Rate
$p_i = \frac{1}{n}, \forall\, i \in [n]$ $\tau_i = \tau, \forall\, i \in [n]$	Sharma et al. (2022) with $P = n$: $\mathcal{O}\left(\frac{1}{(\tau PT)^{1/4}}\right) + \mathcal{O}\left(\frac{(n\tau)^{3/2}}{\sqrt{T}}\right)$ **Ours** with $P < n$: $\begin{cases} \mathcal{O}\left(\sqrt{\sigma_G}\left(\frac{n-P}{n-1}\frac{1}{PT}\right)^{1/4} + \sqrt{\sigma_L}\left(\frac{1}{\tau PT}\right)^{1/4}\right) + \\ \mathcal{O}\left(\frac{(\tau P)^{1/4}}{T^{3/4}}\left(1 + \tau\frac{n-P}{n-1}\right)^{-1/4}\right) + \mathcal{O}\left(\frac{1}{T^{3/4}}\left[\frac{\sigma_L^2}{\tau} + (G_{\mathbf{x}}^2 + \sigma_G^2)\right]\right) \end{cases}$
Ours: $p_i \neq p_j,\ \tau_i \neq \tau_j$ $\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i$	$\mathcal{O}\left(\left(\sqrt{\sigma_G}\left(\frac{n-P}{n-1}\frac{n\max_i p_i}{PT}\right)^{1/4} + \sqrt{\sigma_L}\left(\frac{n\tau_{\text{eff}}}{\tau_{\text{eff}}PT}\sum_{i=1}^n \frac{p_i^2}{\tau_i}\right)^{1/4}\right)\sqrt[8]{1 + \frac{n}{P}\left\|\mathbf{p}\right\|_2^2}\right)$ $+\mathcal{O}\left(\frac{(\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}\left(1 + \frac{n}{P}\left\|\mathbf{p}\right\|_2^2\right)\left(n\tau_{\text{eff}}\sum_{i=1}^n \frac{p_i^2}{\tau_i} + \tau_{\text{eff}}\frac{n-P}{n-1}n\max_i p_i\right)^{-1/4}\right)$ $+\mathcal{O}\left(\frac{1}{T^{3/4}}\left[\sigma_L^2\frac{\sum_{i=1}^n p_i\tau_i}{\bar{\tau}^2} + (G_{\mathbf{x}}^2 + \sigma_G^2)\max_i \frac{\tau_i^2}{\bar{\tau}^2}\right]\right)$

- **Partial Client Participation Error:** $\mathcal{O}(\frac{\sigma_G}{\sqrt{PT}})$ is the *most significant* component of convergence error. Unlike the other two errors, it does not decrease with increasing local updates τ_{eff}. Consequently, we do not observe communication savings by performing multiple local updates at the clients. It remains an open problem to achieve speedup in terms of local updates in partial participation settings.
- **Unequal client weights:** if the clients are weighted disparately, we observe an increase in the stochastic gradient complexity. To see this, let $\tau_i \equiv \tau$. The resulting bound is $\mathcal{O}\Big(\sigma_G\sqrt{\frac{(n-P)n\|\mathbf{p}\|_\infty}{(n-1)PT}} + \frac{\sigma_L\sqrt{n}\|\mathbf{p}\|_2}{\sqrt{P\tau_{\text{eff}}T}} + \frac{1}{T}[\frac{\sigma_L^2}{\bar{\tau}} + \sigma_G^2]\Big)$. Since $\|\mathbf{p}\|_\infty, \|\mathbf{p}\|_2 \leq 1$, in the worst case (when only one of the clients has all the weight), the complexity is worse by a factor of n. This happens because the client sampling is not done in proportion to their weights. Rather, the server first samples the clients uniformly, and then scales their updates to get an unbiased estimator (3). We leave exploring non-uniform WOR sampling further as a future direction.

Corollary 1.2 (Improved Communication Cost)**.** *Suppose all the clients are weighted equally ($p_i = 1/n$ for all i), with each carrying out τ local steps of SGDA. Further, assume Φ is bounded from below. Then, to reach $\mathbf{x}$ such that $\mathbb{E}\|\nabla\Phi(\mathbf{x})\| \leq \epsilon$,*

- *Under full participation, the per-client gradient complexity of Fed-Norm-SGDA is $T\tau = \mathcal{O}(\kappa^4/(n\epsilon^4))$. The number of communication rounds required is $T = \mathcal{O}(\kappa^2/\epsilon^2)$.*
- *Under partial participation, the per-client gradient complexity of Fed-Norm-SGDA is $\mathcal{O}(\kappa^4/(P\epsilon^4))$. In general, running multiple local updates does not yield any communication savings. However, in the special case when inter-client data heterogeneity $\sigma_G = 0$, the communication cost is $\mathcal{O}(\kappa^2/\epsilon^2)$.*

Remark 2. The gradient complexity in Corollary 1.2 is optimal in ϵ, and achieves linear speedup in the number of participating clients. The communication complexity improves the corresponding results in Deng & Mahdavi (2021); Sharma et al. (2022). We match the communication cost in the recent work Yang et al. (2022a). In addition, our work considers a more general FL setting with unequally weighted clients ($p_i \neq p_j$), running unequal local updates ($\tau_i \neq \tau_j$), using distinct local solvers ($\boldsymbol{a}_i \neq \boldsymbol{a}_j$).

Extending the Results to Nonconvex-PL Functions

Assumption 6. A function f satisfies μ-PL condition in $\mathbf{y}$ ($\mu > 0$), if for any fixed $\mathbf{x}$: 1) $\max_{\mathbf{y}'} f(\mathbf{x}, \mathbf{y}')$ has a nonempty solution set; and 2) for all $\mathbf{y}$

$$\|\nabla_y f(\mathbf{x}, \mathbf{y})\|^2 \geq 2\mu(\max_{\mathbf{y}'} f(\mathbf{x}, \mathbf{y}') - f(\mathbf{x}, \mathbf{y})).$$

Remark 3. If Assumptions 1, 2, 3, 4 hold, and the global function F satisfies Assumption 6, then for appropriately chosen learning rates (Appendix B.5), the bound in Theorem 1 holds.

5.2 Non-convex-Concave (NC-C) Case

In this subsection, we consider smooth nonconvex functions which satisfy the following assumptions.

Assumption 7 (Concavity)**.** The function f is concave in $\mathbf{y}$ if for a fixed $\mathbf{x} \in \mathbb{R}^{d_1}$, for all $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{d_2}$,

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

Assumption 8 (Lipschitz continuity in $\mathbf{x}$)**.** Given a function f, there exists a constant $G_{\mathbf{x}}$, such that for each $\mathbf{y} \in \mathbb{R}^{d_2}$, and all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_1}$,

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

The envelope function $\Phi(\mathbf{x}) = \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ used so far, may no longer be smooth in the absence of a unique maximizer. However, $\Phi(\cdot)$ is weakly convex (Lin et al., 2020a, Lemma 4.7). Therefore, we use the alternate definition of stationarity, proposed in Davis & Drusvyatskiy (2019), utilizing the Moreau envelope of Φ.

Definition 4 (Moreau Envelope)**.** The function ϕ_λ is the λ-Moreau envelope of ϕ, for $\lambda > 0$, if for all $\mathbf{x} \in \mathbb{R}^{d_x}$,

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

Drusvyatskiy & Paquette (2019) showed that a small $\|\nabla \phi_\lambda(\mathbf{x})\|$ indicates the existence of some point $\widetilde{\mathbf{x}}$ in the vicinity of $\mathbf{x}$, that is *nearly stationary* for ϕ. Hence, in our case, we focus on minimizing $\|\nabla \Phi_\lambda(\mathbf{x})\|$.

Proposed Algorithm. For nonconvex-concave functions, we use Fed-Norm-SGDA+. The $\mathbf{x}$-updates are identical to Fed-Norm-SGDA. For the $\mathbf{y}$ updates however, the clients compute stochastic gradients $\nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ keeping the x-component fixed at $\widehat{\mathbf{x}}^{(s)}$ for S communication rounds. This *trick*, originally proposed in Deng & Mahdavi (2021), gives the analytical benefit of a double-loop algorithm (which updates $\mathbf{y}$ several times before updating $\mathbf{x}$ once) while also updating $\mathbf{x}$ simultaneously.

Theorem 2. *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 2, 3, 4, 7, 8, the $\mathbf{y}$ iterates are bounded, and the server selects $|\mathcal{C}^{(t)}| = P$ clients for all t. With appropriate client and server learning rates, (η_x^c, η_y^c) and (γ_x^s, γ_y^s) respectively (see Appendix C.2), the iterates of Fed-Norm-SGDA+ satisfy*

$$\begin{aligned}\min_{t\in[T]} \mathbb{E}\left\|\nabla \widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq & \underbrace{\mathcal{O}\left(\left(\sigma_G^2 \frac{n-P}{n-1} \bar{\Delta}_{\widetilde{\Phi}} \frac{E_w}{PT} \sqrt{1+F_w}\right)^{1/4}\right)}_{\textit{Partial participation error}} + \underbrace{\mathcal{O}\left(\frac{C_w \sigma_L^2 + D(G_{\mathbf{x}}^2 + \sigma_G^2)}{\bar{\tau}^2 T^{3/4}}\right)}_{\textit{Local updates error}} \\ & + \underbrace{\mathcal{O}\left(\left(\frac{\bar{\Delta}_{\widetilde{\Phi}} \sigma_L^2 A_w}{\tau_{\mathit{eff}} PT} \sqrt{1+F_w}\right)^{1/4} + \frac{\bar{\Delta}_{\widetilde{\Phi}}(1+F_w)}{T^{3/4}} \left(\frac{\tau_{\mathit{eff}} P}{A_w + \tau_{\mathit{eff}} \frac{n-P}{n-1} E_w}\right)^{1/4}\right)}_{\textit{Error with full synchronization}},\end{aligned} \tag{6}$$

where $\Phi_{1/2L_f}$ *is the Moreau envelope of* Φ*, and* $\bar{\Delta}_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}_{1/2L_f}(\mathbf{x}_0) - \min_{\mathbf{x}} \widetilde{\Phi}_{1/2L_f}(\mathbf{x})$*. The constants* $A_w, C_w, D, E_w, \bar{\tau}, \tau_{eff}$ *are defined in Theorem 1, and* $F_w \triangleq \frac{n(n-P)}{P(n-1)} \sum_{i=1}^n w_i^2$*.*

See Appendix C for the proof. Theorem 2 states convergence for a surrogate objective $\widetilde{F}$. Next, we see convergence for the true objective F.

Corollary 2.1 (Convergence in terms of F)**.** *Given envelope functions* $\Phi(\mathbf{x}) \triangleq \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$*,* $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x}, \mathbf{y})$*, under the conditions of Theorem 2,*

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

where $\Phi_{1/2L_f}$ *is the Moreau envelope of* Φ*,* $\widetilde{\mathbf{x}}^{(t)} \triangleq \arg\min_{\mathbf{x}'} \{\widetilde{\Phi}(\mathbf{x}') + L_f \left\|\mathbf{x}' - \mathbf{x}^{(t)}\right\|^2\}$*,* $\bar{\mathbf{x}}^{(t)} \triangleq \arg\min_{\mathbf{x}'} \{\Phi(\mathbf{x}') + L_f \left\|\mathbf{x}' - \mathbf{x}^{(t)}\right\|^2\}$*, for all* t*,* ϵ'_{opt} *is the error bound in (6).*

Similar to Corollary 1.1, if we replace $\{w_i\}$ with $\{p_i\}$ for all $i \in [n]$ in the server updates in Algorithm 1, then $\widetilde{F} \equiv F$, and $\widetilde{\mathbf{x}}^{(t)}$ and $\bar{\mathbf{x}}^{(t)}$ are identical for all t. Consequently, Theorem 2 gives us convergence in terms of the true objective F. For the rest of this subsection, we assume $w_i = p_i$ for all $i \in [n]$.

Remark 4. Some existing works do not require Assumption 8 for NC-C functions, and also improve the convergence rate. However, these methods either have a double-loop structure Rafique et al. (2021); Zhang et al. (2022), or work with deterministic problems Xu et al. (2020); Zhang et al. (2020). Proposing a single-loop method for stochastic NC-C problems with the same advantages is an open problem.

Again, in Table 2, we specialize the bound in (6) to SGDA+ based local updates. As in the last section

- Partial client participation is the *most significant* source of convergence error.
- Unequal client weights ($p_i \neq p_j$) can increase the stochastic gradient complexity, due to the presence of $n \left\|\mathbf{p}\right\|_\infty, n \left\|\mathbf{p}\right\|_2^2$ factors.

Corollary 2.2 (Improved Communication Cost)**.** *Suppose all the clients are weighted equally ($p_i = 1/n$ for all i), with each carrying out τ local steps of SGDA+. Further, assume that* $\Phi_{1/2L_f}$ *is bounded from below. Then, to reach* $\mathbf{x}$ *such that* $\mathbb{E}\|\nabla \Phi_{1/2L_f}(\mathbf{x})\| \leq \epsilon$*,*

- *Under full participation, the per-client gradient complexity of Fed-Norm-SGDA+ is* $T\tau = \mathcal{O}(1/(n\epsilon^8))$*. The number of communication rounds required is* $T = \mathcal{O}(1/\epsilon^4)$*.*
- *Under partial participation, the per-client gradient complexity of Fed-Norm-SGDA+ is* $\mathcal{O}(1/(P\epsilon^8))$*. In general, running multiple local updates does not yield any communication savings. However, in the special case when inter-client data heterogeneity* $\sigma_G = 0$*, the communication cost is* $\mathcal{O}(1/\epsilon^4)$*.*

In terms of communication requirements, we achieve massive savings (compared to $\mathcal{O}(1/\epsilon^7)$ in Sharma et al. (2022)). Our gradient complexity results achieve linear speedup in the number of participating clients. Further, as stated earlier, our work considers a more general FL setting with unequally weighted clients ($p_i \neq p_j$), running unequal local updates ($\tau_i \neq \tau_j$), using distinct local solvers ($\boldsymbol{a}_i \neq \boldsymbol{a}_j$).

Extending the Results to Nonconvex-One-Point-Concave Functions. One-point-convexity has been observed in SGD dynamics during neural network training Li & Yuan (2017); Kleinberg et al. (2018).

Assumption 9 (One-point-Concavity in $\mathbf{y}$)**.** The function f is said to be one-point-concave in $\mathbf{y}$ if fixing $\mathbf{x} \in \mathbb{R}^{d_1}$, for all $\mathbf{y} \in \mathbb{R}^{d_2}$,

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

where $\mathbf{y}^*(\mathbf{x}) \in \arg\max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$.

It turns out, Theorem 2 holds for the more general class of nonconvex-one-point-concave (NC-1PC) functions. See Appendix C.4 for more details.

Remark 5. Suppose Assumptions 1, 3, 2, 4, 8 hold. Suppose for all $\mathbf{x}$, all the f_i's satisfy Assumption 9 at a common global minimizer $\mathbf{y}^*(\mathbf{x})$. Then, the bound in Theorem 2 holds.

Remark 6. Hence, we settle the conjecture posed in Sharma et al. (2022) that *linear speedup* can be achieved for NC-1PC functions. As an intermediate step in our proof, we show convergence of Local SGD for one-point-convex functions. This extends the convex result for Local SGD to a larger class of functions.

6 Experiments

In this section, we evaluate the empirical performance of the proposed algorithms. We consider a robust neural training problem Sinha et al. (2017); Nouiehed et al. (2019), and a fair classification problem Mohri et al. (2019); Deng et al. (2020). Due to space constraints, additional details of our experiments, and some additional results are included in Appendix D. Our experiments were run on a network of $n = 15$ clients, each equipped with an NVIDIA TitanX GPU. We model data heterogeneity across clients using Dirichlet distribution Wang et al. (2019) with parameter α, $\text{Dir}_n(\alpha)$. Small $\alpha \Rightarrow$ higher heterogeneity across clients.

Robust NN training. We consider the following robust neural network (NN) training problem.

$$\min_{\mathbf{x}} \max_{\|\mathbf{y}\|^2 \leq 1} \sum_{j=1}^{N} \ell\left(h_{\mathbf{x}}(\mathbf{a}_i + \mathbf{y}), b_i\right), \tag{7}$$

where $\mathbf{x}$ denotes the NN parameters, (a_i, b_i) denote the feature and label of the i-th sample, $\mathbf{y}$ denotes the adversarially added feature perturbation, and $h_{\mathbf{x}}$ denotes the NN output.

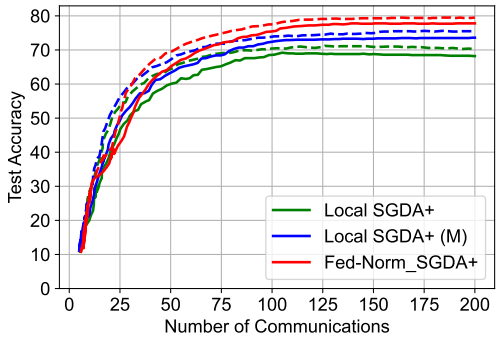

Figure 3: Comparison of the effect of heterogeneous number of local updates $\{\tau_i\}$ on the performance of Fed-Norm-SGDA+ (Algorithm 1), Local SGDA+, and Local SGDA+ with momentum, while solving (7) on CIFAR10 dataset, with VGG11 model. The solid (dashed) curves are for $E = 5$ ($E = 7$), and $\alpha = 0.1$.

Impact of system heterogeneity. In Figure 3, we compare the effect of heterogeneous number of local updates across clients, on the performance of our proposed Fed-Norm-SGDA+. We compare with Local SGDA+ Deng & Mahdavi (2021), and Local SGDA+ with momentum Sharma et al. (2022). Clients sample the number of epochs they run locally via uniform sampling over the set $\{2 \ldots, E\}$, i.e., $\tau_i \sim Unif[2 : E]$. We observe that Fed-Norm-SGDA+ adapts well to system heterogeneity and outperforms both existing methods.

Impact of partial participation and heterogeneity. Next, we compare the impact of different levels of partial client participation on performance. We compare the full participation setting ($n = 15$) with $P = 5, 10$. Clients sample the number of epochs they run locally via $\tau_i \sim Unif[2, 5]$. We plot the results for two different values of the data heterogeneity parameter $\alpha = 0.1, 1.0$. As seen in all our theoretical results where partial participation was the most significant component of convergence error, smaller values of P result in performance loss. Further, higher inter-client heterogeneity (modeled by smaller values of α) results in worse performance. We further explore the impact of α on performance in Appendix D.

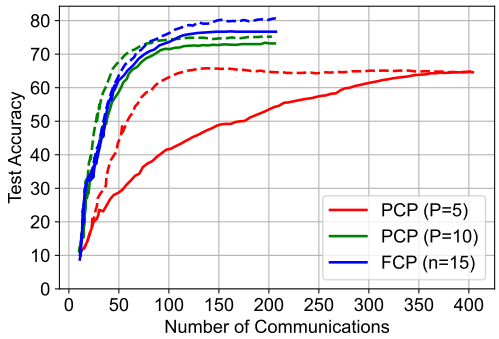

Figure 4: Comparison of the effects of partial client participation (PCP) on the performance of Fed-Norm-SGDA+, for the robust NN training problem on the CIFAR10 dataset, with the VGG11 model. The figure shows the robust test accuracy. The solid (dashed) curves are for $\alpha = 0.1$ ($\alpha = 1.0$).

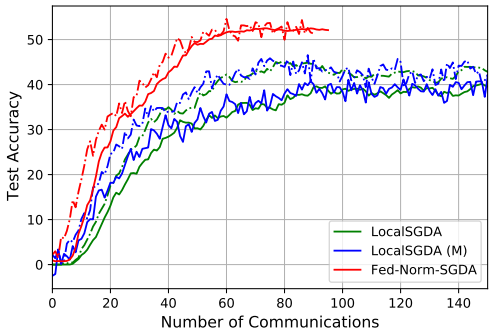

Figure 5: Comparison of Local SGDA, Local SGDA with momentum, and Fed-Norm-SGDA, for the fair classification task on the CIFAR10 dataset, with the VGG11 model. The solid (dashed) curves are for $E = 5$ ($E = 7$), $\alpha = 0.1$.

Fair Classification. We consider minimax formulation of the fair classification problem Mohri et al. (2019); Nouiehed et al. (2019).

$$\min_{\mathbf{x}} \max_{\mathbf{y} \in \Delta_C} \sum_{c=1}^{C} y_c F_c(\mathbf{x}) - \frac{\lambda}{2} \|\mathbf{y}\|^2 , \tag{8}$$

where $\mathbf{x}$ denotes the parameters of the NN, $\{F_c\}_{c=1}^{C}$ denote the loss corresponding to class c, and Δ_C is the C-dimensional probability simplex. In Figure 5, we plot the worst distribution test accuracy achieved by Fed-Norm-SGDA, Local SGDA Deng & Mahdavi (2021) and Local SGDA with momentum Sharma et al. (2022). As in Figure 3, clients sample $\tau_i \sim Unif[2, E]$. We plot the test accuracy on the worst distribution in each case. Again, Fed-Norm-SGDA outperforms existing methods.

7 Conclusion

In this work, we considered nonconvex minimax problems in the federated setting, where in addition to inter-client data heterogeneity and partial client participation, there is system heterogeneity as well. Clients may run unequal number of local update steps, using different local solvers. In such settings, we observed that existing methods, such as Local SGDA, might converge to the stationary point of an objective quite different from the original intended objective. We showed that normalizing individual client contributions solves this problem. Using our generalized framework, we analyzed several classes of nonconvex minimax

functions and significantly improved existing computation and communication complexity results. Potential future directions include analyzing federated systems with unpredictable client presence Yang et al. (2022b).

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. *arXiv preprint arXiv:2110.03313*, 2021a.

Aleksandr Beznosikov, Alexander Rogozin, Dmitry Kovalev, and Alexander Gasnikov. Near-optimal decentralized algorithms for saddle point problems over time-varying networks. In *International Conference on Optimization and Applications*, pp. 246–257. Springer, 2021b.

Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, and Alexander Gasnikov. Distributed saddle-point problems under similarity. In *Advances in Neural Information Processing Systems*, volume 34, 2021c.

Aleksandr Beznosikov, Vadim Sushko, Abdurakhmon Sadiev, and Alexander Gasnikov. Decentralized personalized federated min-max problems. *arXiv preprint arXiv:2106.07289*, 2021d.

Mingzhe Chen, Nir Shlezinger, H Vincent Poor, Yonina C Eldar, and Shuguang Cui. Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*, 118(17):e2024789118, 2021.

Ziyi Chen, Yi Zhou, Tengyu Xu, and Yingbin Liang. Proximal gradient descent-ascent: Variable convergence under kł geometry. In *International Conference on Learning Representations*, 2020.

Hanseul Cho and Chulhee Yun. Sgda with shuffling: faster convergence for nonconvex-p {\L} minimax optimization. *arXiv preprint arXiv:2210.05995*, 2022.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pp. 1387–1395. PMLR, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15111–15122, 2020.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, volume 27, 2014.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32:11082–11094, 2019.

Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, pp. 3973–3983. PMLR, 2020.

Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*, 2021.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pp. 4880–4889. PMLR, 2020.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pp. 2698–2707. PMLR, 2018.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pp. 6175–6186. PMLR, 2021.

Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Luofeng Liao, Li Shen, Jia Duan, Mladen Kolar, and Dacheng Tao. Local adagrad-type algorithm for stochastic convex-concave minimax problems. *arXiv preprint arXiv:2106.10022*, 2021.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020a.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pp. 2738–2779. PMLR, 2020b.

Weijie Liu, Aryan Mokhtari, Asuman Ozdaglar, Sarath Pattathil, Zebang Shen, and Nenggan Zheng. A decentralized proximal point-type method for saddle point problems. *arXiv preprint arXiv:1910.14380*, 2019.

Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

Luo Luo and Cheng Chen. Finding second-order stationary point for nonconvex-strongly-concave minimax problem. *arXiv preprint arXiv:2110.04814*, 2021.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20566–20577, 2020.

Luo Luo, Guangzeng Xie, Tong Zhang, and Zhihua Zhang. Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*, 2021.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Dmitriy Metelev, Alexander Rogozin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized saddle-point problems with different constants of strong convexity and strong concavity. *arXiv preprint arXiv:2206.00090*, 2022.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34: 14606–14619, 2021.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171 (1):115–166, 2018.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14934–14942, 2019.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear TD learning. *arXiv preprint arXiv:2008.10103*, 2020.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pp. 1–35, 2021.

Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21554–21565, 2020.

Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity. *IEEE Journal on Selected Areas in Information Theory*, 2022.

Alexander Rogozin, Alexander Beznosikov, Darina Dvinskikh, Dmitry Kovalev, Pavel Dvurechensky, and Alexander Gasnikov. Decentralized distributed optimization for saddle point problems. *arXiv preprint arXiv:2102.07758*, 2021.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pp. 19683–19730. PMLR, 2022.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2017.

Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

Zhenyu Sun and Ermin Wei. A communication-efficient algorithm with linear convergence for federated minimax learning. *arXiv preprint arXiv:2206.01132*, 2022.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Quoc Tran-Dinh, Deyi Liu, and Lam M Nguyen. Hybrid variance-reduced sgd algorithms for minimax problems with nonconvex-linear function. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11096–11107, 2020.

Isidoros Tziotis, Zebang Shen, Ramtin Pedarsani, Hamed Hassani, and Aryan Mokhtari. Straggler-resilient personalized federated learning. *arXiv preprint arXiv:2206.02078*, 2022.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2019.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of local-update sgd algorithms. *Journal of Machine Learning Research*, 22(213):1–50, 2021.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4800–4810, 2020.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

Guangzeng Xie, Luo Luo, Yijiang Lian, and Zhihua Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *International Conference on Machine Learning*, pp. 10504–10513. PMLR, 2020.

Jiahao Xie, Chao Zhang, Yunsong Zhang, Zebang Shen, and Hui Qian. A federated learning framework for nonconvex-pl minimax problems. *arXiv preprint arXiv:2105.14216*, 2021.

Eric P Xing, Qirong Ho, Pengtao Xie, and Dai Wei. Strategies and principles of distributed machine learning on big data. *Engineering*, 2(2):179–195, 2016.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2020.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. In *International Conference on Learning Representations*, 2021.

Haibo Yang, Zhuqing Liu, Xin Zhang, and Jia Liu. Sagda: Achieving $\mathcal{O}(\epsilon^{-2})$ communication complexity in federated min-max learning. *arXiv preprint arXiv:2210.00611*, 2022a.

Haibo Yang, Xin Zhang, Prashant Khanduri, and Jia Liu. Anarchic federated learning. In *International Conference on Machine Learning*, pp. 25331–25363. PMLR, 2022b.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5667–5678, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pp. 5485–5517. PMLR, 2022c.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with o (1/kˆ2) rate on squared gradient norm. In *International Conference on Machine Learning*, pp. 12098–12109. PMLR, 2021.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *International Conference on Machine Learning*, pp. 7184–7193. PMLR, 2019.

Chulhee Yun, Shashank Rajput, and Suvrit Sra. Minibatch vs local sgd with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations*, 2022.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7377–7389, 2020.

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 482–492. PMLR, 2021.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *arXiv preprint arXiv:2205.15084*, 2022.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Single-client minimax . . . 3

2.2 Distributed/Federated Minimax . . . 3

3 Preliminaries **4**

4 Algorithm for Heterogeneous Federated Minimax Optimization **5**

4.1 Limitations of Local SGDA . . . 5

4.2 Proposed Normalized Federated Minimax Algorithm . . . 6

5 Convergence Results **7**

5.1 Non-convex-Strongly-Concave (NC-SC) Case . . . 8

5.2 Non-convex-Concave (NC-C) Case . . . 10

6 Experiments **12**

7 Conclusion **13**

A Background **20**

A.1 Gradient Aggregation with Different Solvers at Clients . . . 20

A.2 Auxiliary Results . . . 21

B Convergence of Fed-Norm-SGDA for Nonconvex-Strongly-Concave Functions (Theorem 1) **22**

B.1 Intermediate Lemmas . . . 23

B.2 Proof of Theorem 1 . . . 24

B.3 Proofs of the Intermediate Lemmas . . . 27

B.4 Auxiliary Lemmas . . . 37

B.5 Convergence under Polyak Łojasiewicz (PL) Condition . . . 38

C Convergence of Fed-Norm-SGDA+ for Nonconvex Concave Functions (Theorem 2) **38**

C.1 Intermediate Lemmas . . . 39

C.2 Proof of Theorem 2 . . . 41

C.3 Proofs of the Intermediate Lemmas . . . 45

C.4 Extending the result for Nonconvex One-Point-Concave (NC-1PC) Functions . . . 50

D Additional Experiments **50**

Appendix

The appendix are organized as follows. In Section A we mention some basic mathematical results and inequalities which are used throughout the paper. In Appendix B we prove the non-asymptotic convergence of Fed-Norm-SGDA (Algorithm 1) for smooth nonconvex-strongly-concave (and nonconvex-PŁ) functions, and derive gradient complexity and communication cost of the algorithm to achieve an ϵ-stationary point. In Appendix C, we prove the non-asymptotic convergence of Fed-Norm-SGDA+ (Algorithm 1) for smooth nonconvex-concave and nonconvex-one-point-concave functions. Finally, in Appendix D we provide the details of the additional experiments we performed.

A Background

A.1 Gradient Aggregation with Different Solvers at Clients

Local SGDA. Suppose $\tau_i^{(t)} = \tau_{\text{eff}}^{(t)} = \tau$ for all $i \in [n]$, $t \in [T]$. Also, $a_i^{(t,k)} = 1$ for all $k \in [\tau], t$. Then, the local iterate updates in Algorithm 1-Fed-Norm-SGDA reduce to (the updates for Fed-Norm-SGDA+ are analogous)

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

for $k \in \{0, \dots, \tau - 1\}$ and the gradient aggregate vectors $(\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)})$ are simply the average of individual gradients

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

Note that these are precisely the iterates of LocalSGDA proposed in Deng & Mahdavi (2021); Sharma et al. (2022), with the only difference that in LocalSGDA, the clients communicate the iterates $\{\mathbf{x}_i^{(t,\tau)}, \mathbf{y}_i^{(t,\tau)}\}$ to the server, which averages them to compute $\{\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}\}$. While here, the clients communicate $\{\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}\}$. Also, in Fed-Norm-SGDA, the clients and server use separate learning rates, which results in tighter bounds on the local-updates error.

With Momentum in Local Updates. Suppose each local client uses a momentum buffer with momentum scale ρ. Then, for $k \in \{0, \dots, \tau_i^{(t)} - 1\}$

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

Simple calculations show that the coefficient of $\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ and $\nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)})$ in the gradient aggregate vectors $(\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)})$ is

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

Therefore, the aggregation vector is $\bar{\boldsymbol{a}}_i^{(t)} = \frac{1}{1-\rho}[1 - \rho^{\tau_i^{(t)}}, 1 - \rho^{\tau_i^{(t)}-1}, \dots, 1 - \rho]$, and

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

A.2 Auxiliary Results

Remark 7 (Impact of heterogeneity σ_G even with $\tau = 1$). Consider two simple minimization problems:

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

(P1) is a simple distributed minimization problem, with n clients, which we solve using synchronous distributed SGD. At iteration t, each client i computes stochastic gradient $\nabla f_i(\mathbf{x}^{(t)}; \xi_i^{(t)})$, and sends it to the server, which averages these, and takes a step in the direction $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}^{(t)}; \xi_i^{(t)})$. On the other hand, **(P1)** is a centralized minimization problem, where at each iteration t, the agent computes a stochastic gradient estimator with batch-size n, $\frac{1}{n}\sum_{i=1}^{n} \nabla f(\mathbf{x}^{(t)}; \xi_i^{(t)})$. We compare the variance of the two global gradient estimators as follows.

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

Since almost all the existing works consider the local variance bound (Assumption 3) with $\beta_L = 0$, the global gradient estimator in both synchronous distributed SGD **(P1)** and single-agent minibatch SGD **(P2)** have the same $\frac{\sigma_L^2}{n}$ variance bound. Therefore, in most existing federated works on minimization Wang et al. (2020); Yang et al. (2021) and minimax problems Sharma et al. (2022), the full synchronization error only depends on the local variance σ_L^2. However, as seen above, for $\beta_L > 0$, this *apparent equivalence* breaks down. Koloskova et al. (2020), which considers similar local variance assumption as ours for minimization problems, also show similar dependence on heterogeneity σ_G.

Lemma A.1 (Young's inequality)**.** *Given two same-dimensional vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, the Euclidean inner product can be bounded as follows:*

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

for every constant $\gamma > 0$.

Lemma A.2 (Strong Concavity)**.** *A function $g : \mathcal{X} \times \mathcal{Y}$ is strongly concave in $\mathbf{y}$, if there exists a constant $\mu > 0$, such that for all $\mathbf{x} \in \mathcal{X}$, and for all $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$, the following inequality holds.*

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

Lemma A.3 (Jensen's inequality)**.** *Given a convex function f and a random variable X, the following holds.*

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

Lemma A.4 (Sum of squares)**.** *For a positive integer K, and a set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_K$, the following holds:*

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

Lemma A.5 (Quadratic growth condition Karimi et al. (2016))**.** *If function g satisfies Assumptions 1, 5, then for all $\mathbf{x}$, the following conditions holds*

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

Lemma A.6. *For L-smooth, convex function g, the following inequality holds*

$$\mathbb{E}\left\|\nabla g(\mathbf{y})-\nabla g(\mathbf{x})\right\|^2 \leq 2L\left[g(\mathbf{y})-g(\mathbf{x})-\nabla g(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\right]. \tag{9}$$

Lemma A.7 (Proposition 6 in Cho & Yun (2022))**.** *For L-smooth function g which is bounded below by g^*, the following inequality holds for all $\mathbf{x}$*

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

B Convergence of Fed-Norm-SGDA for Nonconvex-Strongly-Concave Functions (Theorem 1)

We organize this section as follows. First, in Appendix B.1 we present some intermediate results, which we use to prove the main theorem. Next, in Appendix B.2, we present the proof of Theorem 1, which is followed by the proofs of the intermediate results in Appendix B.3. Appendix B.4 contains some auxiliary results. Finally, in Appendix B.5 we discuss the convergence result for nonconvex-PL functions.

The problem we solve is

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

We define $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$ and $\widetilde{\mathbf{y}}^*(\mathbf{x}) \in \arg\max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$. Since $\widetilde{F}(\mathbf{x},\cdot)$ is μ-strongly concave, $\widetilde{\mathbf{y}}^*(\mathbf{x})$ is unique. In Fed-Norm-SGDA (Algorithm 1), the client updates are given by

$$\begin{aligned}
\mathbf{x}_i^{(t,k)} &= \mathbf{x}^{(t)} - \eta_x^c \sum_{j=0}^{k-1} a_i^{(j)}(k)\nabla_x f_i(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}),\\
\mathbf{y}_i^{(t,k)} &= \mathbf{y}^{(t)} + \eta_y^c \sum_{j=0}^{k-1} a_i^{(j)}(k)\nabla_y f_i(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}),
\end{aligned} \tag{11}$$

where $1 \leq k \leq \tau_i$. These client updates are then aggregated to compute $\{\mathbf{g}_{\mathbf{x},i}^{(t)}, \mathbf{g}_{\mathbf{y},i}^{(t)}\}$

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

Remark 8. Note that we have made explicit, the dependence on k in $a_i^{(j)}(k)$ above. This was omitted in the main paper to avoid tedious notation. However, for some local optimizers, such as local momentum at the clients (Appendix A.1), the coefficients $a_i^{(j)}(k)$ change with k. We assume in our subsequent analysis that $a_i^{(j)}(k) \leq \alpha$ for all $j \in \{0,1,\ldots,k-1\}$ and for all $k \in \{1,2,\ldots,\tau_i\}$. We also use the notation $\|\boldsymbol{a}_i\| \triangleq \left\|\boldsymbol{a}_i(\tau_i)\right\|$.

At iteration t, the server samples $|\mathcal{C}^{(t)}|$ clients *without* replacement **(WOR)** uniformly at random. While aggregating at the server, client i update is weighed by $\tilde{w}_i = w_i n/|\mathcal{C}^{(t)}|$. The aggregates $(\mathbf{g}_{\mathbf{x}}^{(t)}, \mathbf{g}_{\mathbf{y}}^{(t)})$ computed at the server are of the form

$$\begin{aligned}
\mathbf{g}_{\mathbf{x}}^{(t)} &= \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}, \quad \text{such that} \quad \mathbb{E}_{\mathcal{C}^{(t)}}[\mathbf{g}_{\mathbf{x}}^{(t)}] = \mathbb{E}_{\mathcal{C}^{(t)}}\Big[\sum_{i=1}^n \mathbb{I}(i\in\mathcal{C}^{(t)})\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t)}\Big] = \sum_{i=1}^n w_i \mathbf{g}_{\mathbf{x},i}^{(t)}\\
\mathbf{g}_{\mathbf{y}}^{(t)} &= \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{y},i}^{(t)}, \quad \text{such that} \quad \mathbb{E}_{\mathcal{C}^{(t)}}[\mathbf{g}_{\mathbf{y}}^{(t)}] = \mathbb{E}_{\mathcal{C}^{(t)}}\Big[\sum_{i=1}^n \mathbb{I}(i\in\mathcal{C}^{(t)})\tilde{w}_i\mathbf{g}_{\mathbf{y},i}^{(t)}\Big] = \sum_{i=1}^n w_i \mathbf{g}_{\mathbf{y},i}^{(t)}
\end{aligned} \tag{12}$$

For simplicity of analysis, unless stated otherwise, we assume that $|\mathcal{C}^{(t)}| = P$ for all t. Finally, server updates the $\mathbf{x}, \mathbf{y}$ variables as

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

We define by $\mathcal{F}(t')$ the σ-algebra generated by $\{\{\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\}_{i,k}\}_{t=0}^{t'-1}$. Throughout, we denote the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}(t)]$ by the shorthand $\mathbb{E}_t[\cdot]$.

B.1 Intermediate Lemmas

We begin with the following result from Nouiehed et al. (2019) about the smoothness of $\widetilde{\Phi}(\cdot)$.

Lemma B.1. *If a function $f(\cdot,\cdot)$ satisfies Assumptions 1, 5 (L_f-smoothness and μ-strong concavity in $\mathbf{y}$), then $\phi(\cdot) \triangleq \max_{\mathbf{y}} f(\cdot, \mathbf{y})$ is L_Φ-smooth with $L_\Phi = \kappa L_f/2 + L_f$, where $\kappa = L_f/\mu$ is the condition number.*

Lemma B.2. *Suppose the local client loss functions $\{f_i\}$ satisfy Assumptions 1, 4, and the stochastic oracles for the local functions satisfy Assumption 3. Suppose the server selects P clients in each round. Then the iterates generated by Fed-Norm-SGDA (Algorithm 1) satisfy*

$$\begin{aligned}
&\mathbb{E}_t \left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2 = \mathbb{E}_t \left\| \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)} \right\|^2 \\
&\leq \frac{n}{P}\left(\frac{P-1}{n-1}\right) \mathbb{E}_t \left\| \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2 + \frac{n}{P}\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)]^2 \left[\sigma_L^2 + \beta_L^2 \mathbb{E}_t \left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)})\right\|^2\right] \\
&\quad + \frac{n(n-P)}{n-1}\left[\frac{2L_f^2}{P}\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + (\max_i w_i)\frac{2}{P}\left(\beta_G^2 \left\|\nabla_x \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right\|^2 + \sigma_G^2\right)\right],
\end{aligned} \tag{13}$$

where, $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \triangleq \mathbb{E}_t\left[\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\|^2 + \|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\|^2\right]$ is the iterate drift for client i, at local iteration k in the t-th communication round.

Lemma B.3. *Suppose the local client loss functions $\{f_i\}$ satisfy Assumptions 1, 4, 5, and the stochastic oracles for the local functions satisfy Assumption 3. Also, the server learning rate γ_x^s satisfies $64\tau_{\text{eff}}\gamma_x^s L_\Phi \beta_L^2 \beta_G^2 \frac{n}{P}(\max_i w_i \|\boldsymbol{a}_i\|_2^2/\|\boldsymbol{a}_i\|_1^2) \leq 1$, $8\tau_{\text{eff}}\gamma_x^s L_\Phi (\max_i w_i)\frac{n}{P}\left(\frac{n-P}{n-1}\right)\max\{8\beta_G^2, 1\} \leq 1$, and $8\tau_{\text{eff}}\gamma_x^s L_\Phi \beta_L^2 \frac{n}{P}(\max_{i,k} w_i a_i^{(k)}(\tau_i)/\|\boldsymbol{a}_i\|_1) \leq 1$. Then the iterates generated by Algorithm 1 satisfy*

$$\begin{aligned}
&\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{\Phi}(\mathbf{x}^{(t)})\right] \leq -\frac{7\tau_{\text{eff}}\gamma_x^s}{16}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1 - \frac{n(P-1)}{P(n-1)}\tau_{\text{eff}}\gamma_x^s L_\Phi\right)\mathbb{E}_t\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\quad + \frac{5}{4}\tau_{\text{eff}}\gamma_x^s L_f^2 \sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + \frac{9\tau_{\text{eff}}\gamma_x^s L_f^2}{4\mu}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] \\
&\quad + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2 \sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1} + 2\beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right].
\end{aligned} \tag{14}$$

Remark. The bound in Equation (14) looks very similar to the corresponding one-step decay bound for simple smooth minimization problems. The major difference is the presence of $\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]$, which quantifies the inaccuracy of $\mathbf{y}^{(t)}$ in solving the *max* problem $\max_{\mathbf{y}} \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y})$. The term $\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$ is the client drift and is bounded in Lemma B.4 below.

Lemma B.4. *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 4, 5, and the stochastic oracles for the local functions satisfy Assumption 3. Further, in Algorithm 1, we choose learning rates η_x^c, η_y^c such that $\max\{\eta_x^c, \eta_y^c\} \leq \frac{1}{2L_f(\max_i \|\boldsymbol{a}_i\|_1)\sqrt{2(1+\beta_L^2)}}$. Then, the iterates $\{\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}\}$ generated by Fed-Norm-SGDA*

(Algorithm 1) satisfy

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

where $M_{\mathbf{a}_{-1}} \triangleq \max_i \left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\right)$.

Lemma B.5. *Suppose the local loss functions* $\{f_i\}$ *satisfy Assumptions 1, 4, 5, and the stochastic oracles for the local functions satisfy Assumption 3. The server learning rates* γ_x^s, γ_y^s *satisfy the following conditions:*

$$\tau_{\mathit{eff}}\gamma_y^s\kappa L_f\beta_G^2\frac{n}{P}\max\left\{\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1}\max_i w_i\right\} \leq \frac{1}{64}, \gamma_x^s \leq \frac{\gamma_y^s}{81\kappa^2},$$

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

The client learning rates η_x^c, η_y^c *satisfy* $\eta_y^c L_f\beta_G \leq \frac{1}{16\sqrt{\kappa M_{\mathbf{a}_{-1}}}}$ *and* η_x^c*:* $\eta_x^c L_f\beta_G \leq \frac{1}{64\kappa\sqrt{M_{\mathbf{a}_{-1}}}}$*, respectively. Then the iterates generated by Fed-Norm-SGDA (Algorithm 1) satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]$$
$$\leq \frac{4\left[\widetilde{\Phi}(\mathbf{x}^{(0)})-\widetilde{F}(\mathbf{x}^{(0)},\mathbf{y}^{(0)})\right]}{\tau_{\mathit{eff}}\gamma_y^s\mu T} + \frac{1}{12\mu\kappa^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + \frac{4\tau_{\mathit{eff}}[\gamma_x^s]^2L_\Phi}{\gamma_y^s\mu}\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^{n}w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2$$
$$+ 8\tau_{\mathit{eff}}\gamma_y^s\kappa\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]$$
$$+ 8\kappa L_f\left([\eta_x^c]^2+[\eta_y^c]^2\right)\left[\sigma_L^2\sum_{i=1}^{n}w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right]. \tag{15}$$

Remark 9. The proof of Lemma B.5 differs from similar results in the existing literature Sharma et al. (2022); Yang et al. (2022a). As in these works, if all the clients are running the same number of local steps ($\tau_i = \tau$, for all i), we can define virtual sequences of average iterates $\mathbf{x}^{(t,k)} = \frac{1}{P}\sum_{i\in\mathcal{C}^{(t)}}\mathbf{x}_i^{(t,k)}$, $\mathbf{y}^{(t,k)} = \frac{1}{P}\sum_{i\in\mathcal{C}^{(t)}}\mathbf{y}_i^{(t,k)}$, for all $k \in [0, \tau-1], t$. Define $\mathcal{F}'(t,k)$ as the σ-algebra generated by

$$\mathcal{F}'(t,k) \triangleq \sigma\left\{\{\{\mathbf{x}_i^{(s,k)},\mathbf{x}_i^{(s,k)}\}_{i,k}\}_{s=0}^{t-1}\bigcup\{\{\mathbf{x}_i^{(t,j)},\mathbf{x}_i^{(t,j)}\}_i\}_{j=0}^{k-1}\right\}.$$

Since, conditioned on $\mathcal{F}'(t,k)$, $\mathbf{x}^{t,k+1} \perp \{\nabla_y f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)};\xi_i^{(t,k)})\}_{i=1}^n$, using $\{\mathbf{x}^{(t,k)},\mathbf{y}^{(t,k)}\}$ considerably simplifies the analysis. However, with $\tau_i \neq \tau_j$, the virtual sequences $\{\mathbf{x}^{(t,k)},\mathbf{y}^{(t,k)}\}$ can no longer be defined for all k. Hence, we need an alternate proof strategy.

B.2 Proof of Theorem 1

For the sake of completeness, we first state the full statement of Theorem 1 here.

Theorem. *Suppose the local loss functions* $\{f_i\}_i$ *satisfy Assumptions 1, 3, 4, 5. Suppose the server selects clients using without-replacement sampling scheme* ***(WOR)****. Also, the server learning rates* γ_x^s, γ_y^s *and the client learning rates* η_x^c, η_y^c *satisfy the conditions specified in Lemma B.5. Then the iterates generated by*

Fed-Norm-SGDA (Algorithm 1) satisfy

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

where $\kappa = L_f/\mu$ *is the condition number,* $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$ *is the envelope function,* $\Delta_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}(\mathbf{x}^{(0)}) - \min_{\mathbf{x}}\widetilde{\Phi}(\mathbf{x})$*,* $A_w \triangleq n\tau_{\mathit{eff}}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$*,* $B_w \triangleq n\tau_{\mathit{eff}}\max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$*,* $C_w \triangleq \sum_{i=1}^n w_i(\|\boldsymbol{a}_i\|_2^2 - [\alpha_i^{(t,\tau_i-1)}]^2)$*,* $D \triangleq \max_i(\beta_L^2\|\mathbf{a}_{i,-1}\|_2^2 + \|\mathbf{a}_{i,-1}\|_1^2)$*, where* $\mathbf{a}_{i,-1} \triangleq [a_i^{(0)}, a_i^{(1)}, \ldots, a_i^{(\tau_i-2)}]^\top$ *for all* i *and* $E_w \triangleq n\max_i w_i$*.*

Using $\gamma_y^s = \Theta\left(\sqrt{\frac{P}{\tau_{\mathit{eff}}L_f T\left[\Delta_{\widetilde{\Phi}} + A_w\sigma_L^2 + \left(B_w\beta_L^2 + \frac{n-P}{n-1}E_w\tau_{\mathit{eff}}\right)\sigma_G^2\right]}}\right)$ *and* $\eta_x^c \leq \eta_y^c = \Theta\left(\frac{1}{L_f\bar{\tau}\sqrt{T}}\right)$*, where* $\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i$ *in the bounds above, we get*

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

Proof. Using Lemma B.3, and substituting in the bound on iterates' drift from Lemma B.4, we can bound

$$\begin{aligned}
\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{\Phi}(\mathbf{x}^{(t)})\right] \leq& -\frac{7\tau_{\text{eff}}\gamma_x^s}{16}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1 - \frac{n(P-1)}{P(n-1)}\tau_{\text{eff}}\gamma_x^s L_\Phi\right)\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&+ \frac{9\tau_{\text{eff}}\gamma_x^s L_f^2}{4\mu}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] \\
&+ \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1} + 2\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \\
&+ \frac{5}{2}\tau_{\text{eff}}\gamma_x^s\left([\eta_x^c]^2 + [\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right] \\
&+ 10\tau_{\text{eff}}\gamma_x^s L_f^2 M_{\mathbf{a}_{-1}}\beta_G^2\left[[\eta_x^c]^2\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + L_f\left(2\kappa[\eta_x^c]^2 + [\eta_y^c]^2\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\right].
\end{aligned} \tag{16}$$

Summing (16) over $t = 0, \ldots, T-1$, substituting the bound on $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]$ from Lemma B.5, and rearranging the terms, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \\
&= \mathcal{O}\left(\frac{\kappa^2\Delta_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_y^s T} + \tau_{\text{eff}}\gamma_y^s L_f\kappa^2\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right]\right) \\
&\quad + \mathcal{O}\left(\kappa^2\left([\eta_x^c]^2 + [\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left(\|\boldsymbol{a}_i\|_2^2 - [a_i^{(t,\tau_i-1)}]^2\right) + \sigma_G^2\max_i\left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\right)\right]\right)
\end{aligned} \tag{17}$$

Consequently, using constants A_w, B_w, C_w, D, E_w, (17) can be simplified to

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

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

which completes the proof. □

Convergence in terms of F

Proof of Corollary 1.1. According to the definition of $F(\mathbf{x})$ and $\widetilde{F}(\mathbf{x})$, we have

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

Taking norm, using L_f-smoothness and applying Cauchy–Schwarz inequality, we get

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

where the last inequality uses Assumption 4. Next, note that

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

Therefore, we obtain

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

where ϵ_{opt} denotes the optimization error in the right hand side of (4) in Theorem 1. □

Proof of Corollary 1.2. If clients are weighted equally ($w_i = p_i = 1/n$ for all i), with each carrying out τ steps of local SGDA, from (4) we get

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

- For full client participation, this reduces to

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

  To reach an ϵ-stationary point, assuming $n\tau \le T$, the per-client gradient complexity is $T\tau = \mathcal{O}\left(\frac{\kappa^4}{n\epsilon^4}\right)$. Since $\tau \le T/n$, the minimum number of communication rounds required is $T = \mathcal{O}\left(\frac{\kappa^2}{\epsilon^2}\right)$.

- For partial participation, $\mathcal{O}\Big(\left(\frac{n-P}{n-1}\right)\sigma_G^2\sqrt{\frac{\tau}{PT}}\Big)$ is the dominant term, and we do not get any convergence benefit of multiple local updates. Consequently, per-gradient client complexity and number of communication rounds are both $T\tau = \mathcal{O}\left(\frac{\kappa^4}{P\epsilon^4}\right)$, for $\tau = \mathcal{O}(1)$. However, if the data across clients comes from identical distributions ($\sigma_G = 0$), then we recover per-client gradient complexity of $\mathcal{O}\left(\frac{\kappa^4}{P\epsilon^4}\right)$, and number of communication rounds $= \mathcal{O}\left(\frac{\kappa^2}{\epsilon^2}\right)$.

□

B.3 Proofs of the Intermediate Lemmas

Proof of Lemma B.2.

$$
\begin{aligned}
\mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 &= \mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\left(\mathbf{g}_{\mathbf{x},i}^{(t)}-\mathbf{h}_{\mathbf{x},i}^{(t)}+\mathbf{h}_{\mathbf{x},i}^{(t)}\right)\right\|^2 \\
&\overset{(a)}{=} \mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\left(\mathbf{g}_{\mathbf{x},i}^{(t)}-\mathbf{h}_{\mathbf{x},i}^{(t)}\right)\right\|^2 + \mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&= \mathbb{E}_t\left[\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i^2\left\|\mathbf{g}_{\mathbf{x},i}^{(t)}-\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right] + \mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \qquad (\because \mathbb{E}_t[\mathbf{g}_{\mathbf{x},i}^{(t)}] = \mathbf{h}_{\mathbf{x},i}^{(t)} \text{ for all clients } i\in\mathcal{C}^{(t)}) \\
&= \frac{n}{P}\sum_{i=1}^{n} w_i^2\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{x},i}^{(t)}-\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \qquad (\because \tilde{w}_i = w_i n/P \text{ and } \mathbb{P}(i\in\mathcal{C}^{(t)}) = P/n) \\
&\le \frac{n}{P}\sum_{i=1}^{n}\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\left[\sigma_L^2+\beta_L^2\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2\right] + \mathbb{E}_t\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2. \qquad (18)
\end{aligned}
$$

Here, (a) follows from the following reasoning.

$$
\begin{aligned}
&\mathbb{E}_t\left[\sum_{i,j\in\mathcal{C}^{(t)}}\tilde{w}_i\tilde{w}_j\left\langle\mathbf{h}_{\mathbf{x},i}^{(t)},\mathbf{g}_{\mathbf{x},j}^{(t)}-\mathbf{h}_{\mathbf{x},j}^{(t)}\right\rangle\right] \\
&= \mathbb{E}_t\left[\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i^2\mathbb{E}\left[\left\langle\mathbf{h}_{\mathbf{x},i}^{(t)},\mathbf{g}_{\mathbf{x},i}^{(t)}-\mathbf{h}_{\mathbf{x},i}^{(t)}\right\rangle \mid \mathcal{F}(t),\mathcal{C}^{(t)}\right] + \sum_{i\neq j}\tilde{w}_i\tilde{w}_j\underbrace{\mathbb{E}\left[\left\langle\mathbf{h}_{\mathbf{x},i}^{(t)},\mathbf{g}_{\mathbf{x},j}^{(t)}-\mathbf{h}_{\mathbf{x},j}^{(t)}\right\rangle \mid \mathcal{F}(t),\mathcal{C}^{(t)}\right]}_{=0}\right] \\
&\qquad\qquad \text{(Assumption 3; independence of stochastic gradients across clients)} \\
&= \mathbb{E}_t\left[\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i^2\mathbb{E}\left[\left\langle\mathbf{h}_{\mathbf{x},i}^{(t)},\mathbf{g}_{\mathbf{x},i}^{(t)}-\mathbf{h}_{\mathbf{x},i}^{(t)}\right\rangle \mid \mathcal{F}(t),\mathcal{C}^{(t)}\right]\right] \\
&= \mathbb{E}_t\Bigg[\sum_{i\in\mathcal{C}^{(t)}}\frac{\tilde{w}_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}\sum_{j=0}^{\tau_i-1}a_i^{(k)}(\tau_i)a_i^{(j)}(\tau_i)\ \mathbb{E}\Big[\Big\langle\nabla_x f_i\left(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)};\xi_i^{(t,k)}\right)-\nabla_x f_i\left(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)}\right), \\
&\qquad\qquad \nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\Big\rangle \mid \mathcal{F}(t),\mathcal{C}^{(t)}\Big]\Bigg] \\
&= \mathbb{E}_t\Bigg[\sum_{i\in\mathcal{C}^{(t)}}\frac{\tilde{w}_i^2}{\|\boldsymbol{a}_i\|_1^2}\Bigg[\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\ \mathbb{E}\Big[\Big\langle\underbrace{\mathbb{E}\left[\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)};\xi_i^{(t,k)})-\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\middle|\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)}\right]}_{=0},
\end{aligned}
$$

$$\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)})\Big\rangle \Big| \mathcal{F}(t), \mathcal{C}^{(t)}\Big]\Big]$$

$$+ 2\sum_{j<k} a_i^{(k)}(\tau_i) a_i^{(j)}(\tau_i) \mathbb{E}\Big[\Big\langle \underbrace{\mathbb{E}\left[\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}) - \nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}) \Big| \mathbf{x}_i^{(t,j)}, \mathbf{y}_i^{(t,j)}\right]}_{=0},$$

$$\nabla_x f_i(\mathbf{x}_i^{(t,j)}, \mathbf{y}_i^{(t,j)})\Big\rangle \Big| \mathcal{F}(t), \mathcal{C}^{(t)}\Big]\Big]$$

$$= 0. \qquad (\because \text{ Assumption 3})$$

The last inequality in (18) follows from Assumption 1 and 3. Further, we can bound the second term in (18) as follows.

$$\begin{aligned}
&\mathbb{E}_t \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}_t \left\| \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2 + \mathbb{E}_t \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 \qquad \text{((WOR) sampling)} \\
&= \mathbb{E}_t \left\| \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2 + \sum_{i=1}^n \mathbb{E}_t \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}_i w_i \right) \left\| \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2 \right] \\
&\qquad + \sum_{i\neq j} \mathbb{E}_t \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}_t \left\| \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2 + \sum_{i=1}^n \mathbb{E}_t \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}_t \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}_j w_i - \mathbb{I}(i\in\mathcal{C}^{(t)}) \tilde{w}_i w_j + w_i w_j \right) \left\langle \mathbf{h}_{\mathbf{x},i}^{(t)}, \mathbf{h}_{\mathbf{x},j}^{(t)} \right\rangle \right] \\
&= \mathbb{E}_t \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}_t \left[w_i^2 \left\| \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2 \right] + \sum_{i\neq j} \mathbb{E}_t \left[w_i w_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}_t \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}_t \left\| \mathbf{h}_{\mathbf{x},i}^{(t)} \right\|^2, \qquad (19)
\end{aligned}$$

Next, we bound the second term in (19).

$$\begin{aligned}
&\sum_{i=1}^n w_i^2 \mathbb{E}_t \left\| \mathbf{h}_{\mathbf{x},i}^{(t)} - \nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) + \nabla_x f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \right\|^2 \\
&\leq 2L_f^2 \sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i - 1} a_i^{(k)}(\tau_i) \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) + 2(\max_i w_i) \left(\beta_G^2 \left\| \nabla_x \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \right\|^2 + \sigma_G^2 \right), \qquad (20)
\end{aligned}$$

using Assumption 4. $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \triangleq \mathbb{E}_t \left[\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\|^2 + \|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\|^2 \right]$. Substituting (20) in (19), and using the resulting bound in (18) we get the bound in (13). □

Proof of Lemma B.3. Since the local functions $\{f_i\}$ satisfy Assumption 5, $F(\mathbf{x}, \cdot)$ is μ-SC for any $\mathbf{x}$. In the proof, we use the quadratic growth property of μ-SC function $F(\mathbf{x}, \cdot)$, i.e., for any given $\mathbf{x}$

$$\frac{\mu}{2} \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|^2 \leq F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) - F(\mathbf{x}, \mathbf{y}), \qquad \text{for all } \mathbf{y}, \qquad (21)$$

where $\mathbf{y}^*(\mathbf{x}) = \arg\max_{\mathbf{y}'} F(\mathbf{x}, \mathbf{y}')$. Using L_Φ-smoothness of $\widetilde{\Phi}(\cdot)$,

$$\begin{aligned}
\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)})\right] &\leq \mathbb{E}_t\widetilde{\Phi}(\mathbf{x}^{(t)}) - \mathbb{E}_t\left\langle \nabla\widetilde{\Phi}(\mathbf{x}^{(t)}), \tau_{\text{eff}}\gamma_x^s \sum_{i\in\mathcal{C}^{(t)}} \tilde{w}_i \mathbf{g}_{\mathbf{x},i}^{(t)}\right\rangle + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2 \\
&= \widetilde{\Phi}(\mathbf{x}^{(t)}) - \tau_{\text{eff}}\gamma_x^s\mathbb{E}_t\left\langle \nabla\widetilde{\Phi}(\mathbf{x}^{(t)}), \sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\rangle + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2 \qquad \text{(using Assumption 3 and (3))} \\
&= \widetilde{\Phi}(\mathbf{x}^{(t)}) - \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left[\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + \mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\right] + \frac{\tau_{\text{eff}}\gamma_x^s}{2}\mathbb{E}_t\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)}) - \sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\quad + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2. \qquad (22)
\end{aligned}$$

The last term is bounded in Lemma B.2. Next, we bound the third term above.

$$\begin{aligned}
&\mathbb{E}_t\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)}) - \sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&= \mathbb{E}_t\left\|\sum_{i=1}^n w_i\left(\nabla_x f_i(\mathbf{x}^{(t)}, \widetilde{\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 \\
&\qquad\qquad \text{(since } \widetilde{\mathbf{y}}^*(\mathbf{x}) = \arg\max_{\mathbf{y}'}\widetilde{F}(\mathbf{x},\mathbf{y}')\text{)} \\
&\leq 2L_f^2\mathbb{E}_t\left\|\widetilde{\mathbf{y}}^*(\mathbf{x}^{(t)}) - \mathbf{y}^{(t)}\right\|^2 + 2\mathbb{E}_t\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 \\
&\qquad\qquad (L_f\text{-smoothness; Young's inequality}) \\
&\leq \frac{4L_f^2}{\mu}\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}_t\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 \\
&\qquad\qquad \text{(Quadratic growth of } \mu\text{-SC functions (21); Jensen's inequality)} \\
&\leq \frac{4L_f^2}{\mu}\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}_t\left[\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 + \left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2\right] \\
&\qquad\qquad (L_f\text{-smoothness}) \\
&= \frac{4L_f^2}{\mu}\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). \qquad (23)
\end{aligned}$$

where $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \triangleq \mathbb{E}_t\left[\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 + \left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2\right]$. Further, the term containing $\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)})\right\|^2$ in (13) is bounded in Lemma B.7. Substituting the bounds from (23), (13) and Lemma B.7 into (22), we get

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

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

where, the coefficients are simplified, using assumptions on the learning rate γ_x^s

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

This finishes the proof. □

Proof of Lemma B.4. We use the client update equations for individual iterates (11). To bound $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$, first we bound a single component term $\mathbb{E}_t\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2$. For $1 \leq k \leq \tau_i$,

$$
\begin{aligned}
\mathbb{E}_t\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 &= [\eta_x^c]^2\mathbb{E}_t\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\left(\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}\right) - \nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right) + \nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right)\right\|^2 \\
&= [\eta_x^c]^2\left[\mathbb{E}_t\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\left(\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}\right) - \nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right)\right\|^2 + \mathbb{E}_t\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2\right] \\
&\qquad\qquad \text{(using unbiasedness in Assumption 3)} \\
&= [\eta_x^c]^2\left[\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\mathbb{E}_t\left\|\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)};\xi_i^{(t,j)}\right) - \nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2 + \mathbb{E}_t\left\|\sum_{j=0}^{k-1}a_i^{(j)}(k)\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2\right] \\
&\leq [\eta_x^c]^2\left[\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\left(\sigma_L^2 + \beta_L^2\mathbb{E}_t\left\|\nabla_x f_i(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)})\right\|^2\right) + \left(\sum_{j=0}^{k-1}a_i^{(j)}(k)\right)\sum_{j=0}^{k-1}a_i^{(j)}(k)\mathbb{E}_t\left\|\nabla_x f_i\left(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)}\right)\right\|^2\right].
\end{aligned}
\tag{24}
$$

where the last inequality follows from Jensen's inequality (Lemma A.3). Next, note that

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2 \le \frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{k=0}^{\tau_i-2}[a_i^{(j)}(k)]^2 = \sum_{k=0}^{\tau_i-2}[a_i^{(j)}(k)]^2 \le \|\boldsymbol{a}_i\|_2^2 - [a_i^{(t,\tau_i-1)}(\tau_i)]^2,\\
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1} a_i^{(j)}(k) \le \frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{k=0}^{\tau_i-2} a_i^{(j)}(k) = \sum_{k=0}^{\tau_i-2} a_i^{(j)}(k) \le \|\boldsymbol{a}_i\|_1 - [a_i^{(t,\tau_i-1)}(\tau_i)],\\
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2 \le \frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{j=0}^{\tau_i-2}[a_i^{(j)}(k)]^2 \le \alpha\cdot\sum_{k=0}^{\tau_i-2} a_i^{(k)}(\tau_i),
\end{aligned} \tag{25}$$

where $a_i^{(j)}(k) \le \alpha$ for all j, k. We define $\|\boldsymbol{a}_{i,-1}\|_2^2 \triangleq \|\boldsymbol{a}_i\|_2^2 - [a_i^{(t,\tau_i-1)}(\tau_i)]^2$, $\|\boldsymbol{a}_{i,-1}\|_1 \triangleq \|\boldsymbol{a}_i\|_1 - [a_i^{(t,\tau_i-1)}(\tau_i)]$ for the sake of brevity. Using (25), we bound the individual terms in (24).

$$\begin{aligned}
&\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\beta_L^2\left\|\nabla_x f_i(\mathbf{x}_i^{(t,j)},\mathbf{y}_i^{(t,j)})\right\|^2\\
&\le 2\beta_L^2\alpha L_f^2\sum_{j=0}^{\tau_i-2}[a_i^{(j)}(k)]\Delta_{\mathbf{x},\mathbf{y}}^{(t,j)} + 2\beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2.
\end{aligned} \tag{26}$$

Similarly,

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

Substituting (26), (27) in (24), we get

$$\begin{aligned}
\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}_t\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 &\le [\eta_x^c]^2\sigma_L^2\|\boldsymbol{a}_{i,-1}\|_2^2 + 2[\eta_x^c]^2 L_f^2\left(\|\boldsymbol{a}_{i,-1}\|_1 + \beta_L^2\alpha\right)\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\\
&\quad + 2[\eta_x^c]^2\left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2\|\boldsymbol{a}_{i,-1}\|_2^2\right)\left\|\nabla_x f_i(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\|^2.
\end{aligned} \tag{28}$$

Similarly, we can bound the $\mathbf{y}$ error

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

Combining the two bounds in (28) and (29), we get

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

$$+ 2\left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2 \|\boldsymbol{a}_{i,-1}\|_2^2\right)\left[[\eta_x^c]^2 \left\|\nabla_x f_i\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2 + [\eta_y^c]^2 \left\|\nabla_y f_i\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2\right]. \tag{30}$$

Define $A_m \triangleq 2L_f^2\left([\eta_x^c]^2 + [\eta_y^c]^2\right)\max_i \|\boldsymbol{a}_i\|_1 \left(\|\boldsymbol{a}_{i,-1}\|_1 + \beta_L^2\alpha\right)$. Rearranging the terms in (30), and taking the weighted sum over agents, we get

$$\begin{aligned}
&L_f^2 \sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i - 1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) \\
&\leq \frac{\left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2 \sigma_L^2}{1 - A_m} \sum_{i=1}^n w_i \|\boldsymbol{a}_{i,-1}\|_2^2 \\
&\quad + \frac{2L_f^2}{1 - A_m} \sum_{i=1}^n w_i \left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2 \|\boldsymbol{a}_{i,-1}\|_2^2\right)\left[[\eta_x^c]^2 \left\|\nabla_x f_i\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2 + [\eta_y^c]^2 \left\|\nabla_y f_i\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2\right] \\
&\leq \frac{\left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2}{1 - A_m} \sigma_L^2 \sum_{i=1}^n w_i \|\boldsymbol{a}_{i,-1}\|_2^2 + \frac{2L_f^2 M_{\mathbf{a}_{-1}}}{1 - A_m}[\eta_x^c]^2 \left(\beta_G^2 \left\|\sum_{i=1}^n w_i \nabla_x f_i\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2 + \sigma_G^2\right) \\
&\quad + \frac{2L_f^2 M_{\mathbf{a}_{-1}}}{1 - A_m}[\eta_y^c]^2 \left(\beta_G^2 \left\|\sum_{i=1}^n w_i \nabla_y f_i\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2 + \sigma_G^2\right).
\end{aligned} \tag{31}$$

where (31) follows from Assumption 4, and we define $M_{\mathbf{a}_{-1}} \triangleq \max_i \left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2 \|\boldsymbol{a}_{i,-1}\|_2^2\right)$. We bounded $\left\|\nabla_x F\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2$ in Lemma B.6. Similarly, we can bound $\left\|\nabla_y F\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2$ as follows.

$$\begin{aligned}
\left\|\nabla_y F\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right)\right\|^2 &= \left\|\nabla_y F\left(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}\right) - \nabla_y F\left(\mathbf{x}^{(t)}, \mathbf{y}^*(\mathbf{x}^{(t)})\right)\right\|^2 \qquad (\because \mathbf{y}^*(\mathbf{x}) = \arg\max_{\mathbf{y}'} F(\mathbf{x}, \mathbf{y}')) \\
&\leq 2L_f\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right].
\end{aligned} \tag{32}$$

using L_f-smoothness and concavity of $F(\mathbf{x}, \cdot)$ (Lemma A.6). Also, for the choice of η_x^c, η_y^c, we get $A_m \leq 1/2$. Consequently, substituting the two bounds in (31), we complete the proof. □

Proof of Lemma B.5. First we see that

$$\begin{aligned}
&\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{F}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})\right] \\
&= \mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{\Phi}(\mathbf{x}^{(t)})\right] + \left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right] + \mathbb{E}_t\left[\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})\right].
\end{aligned} \tag{33}$$

$\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{\Phi}(\mathbf{x}^{(t)})\right]$ is already bounded in Lemma B.3. We bound $\mathbb{E}_t\left[\widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})\right]$ as follows. Using the notation $\mathbf{z}^{(t)} = (\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ and using L_f-smoothness (Assumption 1) of $\widetilde{F}(\cdot, \cdot)$,

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

Rearranging the terms a bit, we get

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

Next, we bound the individual terms in (34). Using quadratic growth of μ-SC functions (Lemma A.5),

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

To bound $\left\|\nabla_y\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2$, we use similar reasoning as in (23).

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

Substituting the bounds in (35) and Lemma B.6 into (34) we get

$$\begin{aligned}
&\mathbb{E}_t\left[\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})\right]\\
&\leq -\tau_{\text{eff}}\mu\gamma_y^s\left(1-\frac{2\kappa^2\gamma_x^s}{\gamma_y^s}\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]-\frac{\tau_{\text{eff}}\gamma_y^s}{2}\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2\\
&\quad+\tau_{\text{eff}}\gamma_x^s\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2+\frac{\tau_{\text{eff}}\gamma_x^s}{2}\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\frac{\tau_{\text{eff}}^2L_f}{2}\left[[\gamma_y^s]^2\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2+[\gamma_x^s]^2\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{x}}^{(t)}\right\|^2\right]\\
&\quad+\frac{\tau_{\text{eff}}\gamma_y^sL_f^2}{2}\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i).
\end{aligned}\tag{36}$$

$\mathbb{E}_t\|\mathbf{g}_{\mathbf{x}}^{(t)}\|^2$ is already bounded in Lemma B.2. We further substitute the bound on $\mathbb{E}_t\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2$ from Lemma B.7 to get

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

$$
\begin{aligned}
&\leq \frac{n(P-1)}{P(n-1)}\mathbb{E}_t\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + \frac{\sigma_L^2 n}{P}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2n\sigma_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\\
&\quad + \left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\frac{2nL_f^2}{P}\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\\
&\quad + \frac{4n\beta_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left\|\nabla_x\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&\quad + \frac{8n\beta_G^2 L_f\kappa}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]. \qquad (37)
\end{aligned}
$$

Similarly, we can bound $\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2$ to get

$$
\begin{aligned}
\mathbb{E}_t\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2 &\leq \frac{n(P-1)}{P(n-1)}\mathbb{E}_t\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2 + \frac{\sigma_L^2 n}{P}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2 n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2 \max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\\
&\quad + \left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\frac{2nL_f^2}{P}\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\\
&\quad + \frac{4\beta_G^2 L_f n}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]. \qquad (38)
\end{aligned}
$$

Substituting (37), (38) in (36), and again substituting the resulting bound in (33), we get

$$
\begin{aligned}
&\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})\right]\\
&\overset{(a)}{\leq}\left(1-\tau_{\text{eff}}\mu\gamma_y^s\left(1-\frac{\kappa^2\gamma_x^s}{\gamma_y^s}\left[2+\frac{9}{4}\right]-\tau_{\text{eff}}\kappa\gamma_y^s\frac{4n\beta_G^2 L_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)\right)\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&\quad + \tau_{\text{eff}}\gamma_x^s\left(\frac{9}{16}+\tau_{\text{eff}}L_f\gamma_x^s\frac{2n\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)\right)\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&\quad + \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]\\
&\quad + \frac{\tau_{\text{eff}}^2 L_f}{2}\left([\gamma_y^s]^2+[\gamma_x^s]^2\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{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]\\
&\quad - \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 - 1 - \frac{n(P-1)}{P(n-1)}\tau_{\text{eff}}\gamma_x^s L_f\right)\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\
&\quad - \frac{\tau_{\text{eff}}\gamma_y^s}{2}\left(1-\frac{n(P-1)}{P(n-1)}\tau_{\text{eff}}\gamma_y^s L_f\right)\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2\\
&\quad + \tau_{\text{eff}}\left(\frac{3\gamma_x^s}{2}+\frac{3\gamma_y^s}{4}\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)\\
&\overset{(b)}{\leq}\left(1-\tau_{\text{eff}}\mu\gamma_y^s\left(1-\frac{\kappa^2\gamma_x^s}{\gamma_y^s}\left[2+\frac{9}{4}\right]-\tau_{\text{eff}}\kappa\gamma_y^s\frac{4n\beta_G^2 L_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)\right)\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&\quad + \tau_{\text{eff}}\gamma_x^s\left(\frac{9}{16}+\tau_{\text{eff}}L_f\gamma_x^s\frac{2n\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)\right)\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2\\
&\quad + \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2}{2}\frac{n(P-1)}{P(n-1)}(L_\Phi+L_f)\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\tau_{\text{eff}}\gamma_x^s}{2}\left(1 - \frac{n(P-1)}{P(n-1)}\tau_{\text{eff}}\frac{[\gamma_y^s]^2}{\gamma_x^s}L_f\right)\mathbb{E}_t\left\|\sum_{i=1}^n w_i \mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2 \\
&+ \frac{\tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi}{2}\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \sigma_G^2\left(2(\max_i w_i)\frac{n-P}{n-1} + 2\beta_L^2\max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \\
&+ \frac{\tau_{\text{eff}}^2 L_f}{2}\left([\gamma_y^s]^2 + [\gamma_x^s]^2\right)\left[\frac{\sigma_L^2 n}{P}\sum_{i=1}^n \frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2n\sigma_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \\
&+ \tau_{\text{eff}}\gamma_y^s\left[2\left([\eta_x^c]^2 + [\eta_y^c]^2\right)L_f^2\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 4L_f^2 M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2 + [\eta_y^c]^2\right)\sigma_G^2\right] \\
&+ \tau_{\text{eff}}\gamma_y^s\left[8L_f^2 M_{\mathbf{a}_{-1}}\beta_G^2[\eta_x^c]^2\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + 8L_f^3 M_{\mathbf{a}_{-1}}\beta_G^2\left(2\kappa[\eta_x^c]^2 + [\eta_y^c]^2\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\right], \qquad (39)
\end{aligned}
$$

where (a) follows since

$$
\tau_{\text{eff}}L_f\gamma_y^s\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_{i,k}\frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1}\right)\frac{n}{P} \le \frac{1}{4},
$$

and (b) follows by using the bound in Lemma B.4. Next, we simplify the coefficients of different terms in (39).

- Coefficient of $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]$ can be simplified to

$$
\begin{aligned}
&\left(1 - \tau_{\text{eff}}\mu\gamma_y^s\left(1 - \frac{\kappa^2\gamma_x^s}{\gamma_y^s}\frac{17}{4} - \tau_{\text{eff}}\kappa\gamma_y^s\frac{4n\beta_G^2 L_f}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right)\right) \\
&\qquad + \tau_{\text{eff}}\gamma_y^s 8L_f^3 M_{\mathbf{a}_{-1}}\beta_G^2\left(2\kappa[\eta_x^c]^2 + [\eta_y^c]^2\right) \\
&\le 1 - \frac{\tau_{\text{eff}}\gamma_y^s\mu}{4}.
\end{aligned}
$$

  using $\gamma_x^s \le \frac{\gamma_y^s}{17\kappa^2}$, $\tau_{\text{eff}}\gamma_y^s\kappa L_f\beta_G^2\frac{n}{P}\max\left\{\beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1}\max_i w_i\right\} \le \frac{1}{64}$, $\kappa L_f\beta_G\eta_x^c \le \frac{1}{16\sqrt{2M_{\mathbf{a}_{-1}}}}$ and $L_f\beta_G\eta_y^c \le \frac{1}{16\sqrt{\kappa M_{\mathbf{a}_{-1}}}}$.

- Coefficient of $\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2$ can be simplified to

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

- Coefficient of $\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2$ can be simplified to

$$
\begin{aligned}
&\tau_{\text{eff}}\gamma_x^s\left(\frac{9}{16} + \tau_{\text{eff}}L_f\gamma_x^s\frac{2n\beta_G^2}{P}\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right) + \tau_{\text{eff}}\gamma_y^s 8L_f^2 M_{\mathbf{a}_{-1}}\beta_G^2[\eta_x^c]^2 \\
&\le \tau_{\text{eff}}\gamma_x^s\left(\frac{9}{16} + \frac{4\gamma_x^s}{\gamma_y^s}\frac{1}{64\kappa}\right) + \frac{\tau_{\text{eff}}\gamma_y^s}{64\kappa^2} \\
&\le \frac{\tau_{\text{eff}}\gamma_y^s}{48\kappa^2}. \qquad \left(\tfrac{\gamma_x^s}{\gamma_y^s} \le \tfrac{1}{81\kappa^2}\right)
\end{aligned}
$$

  using $\gamma_x^s \le \frac{\gamma_y^s}{81\kappa^2}$, $\eta_x^c L_f\beta_G \le \frac{1}{64\kappa\sqrt{M_{\mathbf{a}_{-1}}}}$ and $\tau_{\text{eff}}\gamma_y^s\kappa L_f\beta_G^2\frac{n}{P}\max\left\{\frac{n-P}{n-1}\max_i w_i, \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right\} \le \frac{1}{64}$.

- Coefficient of σ_L^2 can be simplified to

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

- Coefficient of σ_G^2 can be simplified to

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

Substituting these simplified coefficients in (39), and rearranging the terms, we get

$$\begin{aligned}
&\mathbb{E}_t\left[\widetilde{\Phi}(\mathbf{x}^{(t+1)}) - \widetilde{F}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})\right] \\
&\le \left(1-\frac{\tau_{\text{eff}}\gamma_y^s\mu}{4}\right)\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] + \frac{\tau_{\text{eff}}\gamma_y^s}{48\kappa^2}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 + \tau_{\text{eff}}^2[\gamma_x^s]^2 L_\Phi\frac{n(P-1)}{P(n-1)}\mathbb{E}_t\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 \\
&\quad + \tau_{\text{eff}}^2\left([\gamma_y^s]^2 L_f + [\gamma_x^s]^2 L_\Phi\right)\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] \\
&\quad + 2\tau_{\text{eff}}\gamma_y^s\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right] .
\end{aligned} \tag{40}$$

Summing both sides of (40) over $t = 0,\ldots,T-1$ and rearranging the terms, we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right] \\
&\le \frac{4}{\tau_{\text{eff}}\gamma_y^s\mu}\left[\frac{\widetilde{\Phi}(\mathbf{x}^{(0)}) - \widetilde{F}(\mathbf{x}^{(0)},\mathbf{y}^{(0)})}{T} - \frac{\mathbb{E}\left(\widetilde{\Phi}(\mathbf{x}^{(T)}) - \widetilde{F}(\mathbf{x}^{(T)},\mathbf{y}^{(T)})\right)}{T}\right] + \frac{1}{12\mu\kappa^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}(\mathbf{x}^{(t)})\right\|^2 \\
&\quad + \frac{4\tau_{\text{eff}}[\gamma_x^s]^2 L_\Phi}{\gamma_y^s\mu}\frac{n(P-1)}{P(n-1)}\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2 + 8\kappa L_f\left([\eta_x^c]^2+[\eta_y^c]^2\right)\left[\sigma_L^2\sum_{i=1}^n w_i\|\boldsymbol{a}_{i,-1}\|_2^2 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right] \\
&\quad + 8\tau_{\text{eff}}\gamma_y^s\kappa\frac{n}{P}\left[\sigma_L^2\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + 2\sigma_G^2\left(\frac{n-P}{n-1}\max_i w_i + \beta_L^2\max_i\frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)\right] ,
\end{aligned} \tag{41}$$

where, we use $[\gamma_y^s]^2 L_f \ge [\gamma_x^s]^2 L_\Phi$. This concludes the proof. □

B.4 Auxiliary Lemmas

Lemma B.6. *If the local client function $f_i(\mathbf{x}, \cdot)$ satisfy Assumptions 1, 5 (L_f-smoothness and μ-strong concavity in $\mathbf{y}$), then the function F satisfies*

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

Proof.

$$\begin{aligned}
\left\|\nabla_x \widetilde{F}(\mathbf{x}, \mathbf{y}\right\|^2 &\leq 2\left\|\nabla \widetilde{\Phi}(\mathbf{x})\right\|^2 + 2\left\|\nabla_x F(\mathbf{x}, \mathbf{y}) - \nabla \widetilde{\Phi}(\mathbf{x})\right\|^2 \\
&\leq 2\left\|\nabla \widetilde{\Phi}(\mathbf{x})\right\|^2 + 2L_f^2\left\|\widetilde{\mathbf{y}}^*(\mathbf{x}) - \mathbf{y}\right\|^2 && (L_f\text{-smoothness (Assumption 1)}) \\
&\leq 2\left\|\nabla \widetilde{\Phi}(\mathbf{x})\right\|^2 + \frac{4L_f^2}{\mu}\left[\widetilde{\Phi}(\mathbf{x}) - \widetilde{F}(\mathbf{x}, \mathbf{y})\right]. && \text{(Assumption 5)}
\end{aligned}$$

□

Lemma B.7. *If the local client function $f_i(\mathbf{x}, \cdot)$ satisfy Assumptions 1, 4 and 5, then the iterates $\{\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\}_{i,(t,k)}$ generated by Algorithm 1 satisfy*

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

Proof.

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

Using Lemma B.6 gives the result. □

Lemma B.8. *If the local client function $f_i(\mathbf{x}, \cdot)$ satisfy Assumptions 1, 4 and 5, then the iterates $\{\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\}_{i,(t,k)}$ generated by Algorithm 1 satisfy*

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

Proof. Following closely the proof of Lemma B.7,

$$\sum_{i=1}^{n} \frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)]^2 \mathbb{E} \left\| \nabla_y f_i(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}) \pm \nabla_y f_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \right\|^2$$

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

(L_f-smoothness)

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

(Assumption 4)

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

where the final inequality follows from smoothness and concavity of F in $\mathbf{y}$. □

B.5 Convergence under Polyak Łojasiewicz (PL) Condition

In case the global function satisfies Assumption 6, the results in this section follow with minor modifications. The crucial difference is that Lemma A.6 no longer holds. Lemma B.2 and Lemma B.3 follow exactly. The statement of Lemma B.4 needs some modification, since we use Lemma A.6 in the proof.

Lemma B.9. *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 4, 6, and the stochastic oracles for the local functions satisfy Assumption 3. Under the conditions of Lemma B.9, the iterates $\{\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}\}$ generated by Algorithm 1 satisfy*

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

where $M_{\mathbf{a}_{-1}} \triangleq \max_i \left(\|\boldsymbol{a}_{i,-1}\|_1^2 + \beta_L^2 \|\boldsymbol{a}_{i,-1}\|_2^2 \right)$.

The bound in Lemma B.8 also changes to

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

The same bound in Lemma B.5 holds, but with more stringent conditions on learning rates, namely $\eta_y^c L_f \beta_G \leq \frac{1}{16\kappa\sqrt{M_{\mathbf{a}_{-1}}}}$ and $\tau_{\text{eff}} \gamma_y^s \kappa L_f \beta_G^2 \frac{1}{P} \max \left\{ \kappa \beta_L^2 \max_i \frac{\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, 1 \right\} \leq \frac{1}{64}$. Consequently, the bounds in Theorem 1 hold, under slightly more stringent conditions on the learning rates.

C Convergence of Fed-Norm-SGDA+ for Nonconvex Concave Functions (Theorem 2)

We organize this section as follows. First, in Appendix C.1 we present some intermediate results, which we use in the proof of Theorem 2. Next, in Appendix C.2, we present the proof of Theorem 2, which is followed by the proofs of the intermediate results in Appendix C.3. Finally, we discuss the extension of our results to nonconvex-one-point-concave functions in Appendix C.4.

The problem we solve is

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

We define $\widetilde{\Phi}(\mathbf{x}) \triangleq \max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$ and $\widetilde{\mathbf{y}}^*(\mathbf{x}) \in \arg\max_{\mathbf{y}} \widetilde{F}(\mathbf{x},\mathbf{y})$. Since $\widetilde{F}(\mathbf{x},\cdot)$ is no longer strongly concave, $\mathbf{y}^*(\mathbf{x})$ need not be unique. In Algorithm 1-Fed-Norm-SGDA+ , the client updates are given by

$$\begin{aligned} \mathbf{x}_i^{(t,k)} &= \mathbf{x}^{(t)} - \eta_x^c \sum_{j=0}^{k-1} a_i^{(j)}(k) \nabla_x f_i(\mathbf{x}_i^{(t,j)}, \mathbf{y}_i^{(t,j)}; \xi_i^{(t,j)}), \\ \mathbf{y}_i^{(t,k)} &= \mathbf{y}^{(t)} + \eta_y^c \sum_{j=0}^{k-1} a_i^{(j)}(k) \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,j)}; \xi_i^{(t,j)}), \end{aligned} \tag{42}$$

where $1 \le k \le \tau_i$. The server updates are given by

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \tau_{\text{eff}} \gamma_x^s \mathbf{g}_{\mathbf{x}}^{(t)}, \qquad \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \tau_{\text{eff}} \gamma_y^s \mathbf{g}_{\mathbf{y}}^{(t)}, \tag{43}$$

where $\mathbf{g}_{\mathbf{x}}^{(t)}, \mathbf{g}_{\mathbf{y}}^{(t)}$ are defined in (3). The normalized (stochastic) gradient vectors are defined as

$$\begin{aligned} \mathbf{g}_{\mathbf{x},i}^{(t)} &= \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_x f_i\left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}\right); \quad \mathbf{h}_{\mathbf{x},i}^{(t)} = \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_x f_i\left(\mathbf{x}_i^{(t,k)}, \mathbf{y}_i^{(t,k)}\right), \\ \mathbf{g}_{\mathbf{y},i}^{(t)} &= \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_y f_i\left(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}; \xi_i^{(t,k)}\right); \quad \mathbf{h}_{\mathbf{y},i}^{(t)} = \frac{1}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \nabla_y f_i\left(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}\right). \end{aligned} \tag{44}$$

C.1 Intermediate Lemmas

As discussed in Section 5.2, we analyze the convergence of the smoothed envelope function $\widetilde{\Phi}_{1/2L_f}$. We begin with a bound on the one-step decay of this function.

Lemma C.1 (One-step decay of Smoothed-Envelope)**.** *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 3, 7, and 8. Then, the iterates generated by Algorithm 1-Fed-Norm-SGDA+ satisfy*

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

where $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) = \mathbb{E}\left[\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\|^2 + \|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\|^2\right]$ is the drift of client $i \in [n]$, at the k-th local step of epoch t.

Between two successive synchronization time instants (for example, $t, t+1$), the clients drift apart due to local descent/ascent steps, resulting in the $\{\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)\}_{i,k}$ terms. Also, $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]$ quantifies the error of the inner maximization. In the subsequent lemmas, we bound both these error terms.

Lemma C.2 (Consensus Error)**.** *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 4, 7, and 8. The stochastic oracles for the local functions satisfy Assumption 3. Further, in Algorithm 1-Fed-Norm-SGDA+, we choose the client learning rate η_y^c such that $\eta_y^c \le \frac{1}{2L_f(\max_i \|\boldsymbol{a}_i\|_1)\sqrt{2\max\{1,\beta_L^2\}}}$. Then, the iterates $\{\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}\}$ generated by Algorithm 1-Fed-Norm-SGDA+ satisfy*

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

$$+ 8[\eta_y^c]^2 L_f^3 M_{\mathbf{a}_{-1}} \beta_G^2 \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right],$$

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

Note that consensus error depends on the difference $\mathbb{E}[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})]$. This is different from the term $\mathbb{E}[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})]$ in Lemma C.1. Since in Algorithm 1-Fed-Norm-SGDA+ , the $\mathbf{x}$-component stays fixed at $\widehat{\mathbf{x}}^{(s)}$ for S communication rounds while updating $\mathbf{y}_i^{(t,k)}$, the difference

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

can be interpreted as the optimization error, when maximizing the concave function $F(\widehat{\mathbf{x}}^{(s)}, \cdot)$ over S communication rounds. Next, we bound this error. The following result essentially extends the analysis of FedNova (Wang et al. (2020)) to concave maximization (analogously, convex minimization) problems. We also generalize the corresponding analyses in Khaled et al. (2020); Koloskova et al. (2020) to heterogeneous local updates.

Lemma C.3 (Local SG updates for Concave Maximization)**.** *Suppose the local functions satisfy Assumptions 1, 3, 4 and 7. Further, let* $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ *for all* t*. We run Algorithm 1-Fed-Norm-SGDA+ with client step-size* η_y^c *such that* $64[\eta_y^c]^2 M_{\mathbf{a}_{-1}} L_f^2 \beta_G^2 \frac{n}{P} \leq 1$*. Further, the server step-size* γ_y^s *satisfies*

$$2\tau_{\text{eff}}\gamma_y^s L_f \frac{n}{P} \max\{\beta_G^2, 1\} \max\left\{\beta_L^2 \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1} \max_i w_i\right\} \leq \frac{1}{8},$$

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

Then the iterates generated by Algorithm 1-Fed-Norm-SGDA+ satisfy

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

where $M_{\mathbf{a}_{-1}} \triangleq \max_i \left(\|\mathbf{a}_{i,-1}\|_1^2 + \beta_L^2 \|\mathbf{a}_{i,-1}\|_2^2\right)$.

Remark 10. It is worth noting that the proof of Lemma C.3 does not require global concavity of local functions. Rather, given $\mathbf{x}$, we only need concavity of local functions $\{f_i\}$ at some point $\mathbf{y}^*(\mathbf{x})$. This is precisely the one-point-concavity assumption (Assumption 9) discussed earlier in Deng & Mahdavi (2021); Sharma et al. (2022). Therefore, Lemma C.3 for a much larger class of functions. Further, the bound in Lemma C.3 improves the corresponding bounds derived in existing work. As we discuss in Appendix C.4, this helps us achieve improve complexity results for nonconvex-one-point-concave (NC-1PC) functions.

Next, we bound the difference $\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)}) - \widetilde{F}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\right]$.

Lemma C.4. *Suppose the local functions satisfy Assumptions 1, 3, 4, 8. Then the iterates generated by Algorithm 1-Fed-Norm-SGDA+ satisfy*

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

C.2 Proof of Theorem 2

For the sake of completeness, we first state the full statement of Theorem 2 here.

Theorem 3. *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 3, 4, 7, 8. Further, let $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ for all t. We run Algorithm 1-Fed-Norm-SGDA+ with client step-size η_y^c such that $64[\eta_y^c]^2 M_{\mathbf{a}_{-1}} L_f^2 \max\{\beta_G^2 \frac{n}{P}, 1\} \leq 1$. Further, the server step-size γ_y^s satisfies*

$$2\tau_{\text{eff}}\gamma_y^s L_f \frac{n}{P} \max\{\beta_G^2, 1\} \max\left\{\beta_L^2 \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}, \frac{n-P}{n-1} \max_i w_i\right\} \leq \frac{1}{8},$$

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

Then the iterates generated by Algorithm 1-Fed-Norm-SGDA+ satisfy

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

where $\widetilde{\Phi}_{1/2L_f}(\mathbf{x}) \triangleq \min_{\mathbf{x}'} \widetilde{\Phi}(\mathbf{x}') + L_f \|\mathbf{x}' - \mathbf{x}\|^2$ is the envelope function, $\bar{\Delta}_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}_{1/2L_f}(\mathbf{x}_0) - \min_{\mathbf{x}} \widetilde{\Phi}_{1/2L_f}(\mathbf{x})$, $A_w \triangleq n\tau_{\text{eff}} \sum_{i=1}^n \frac{w_i^2 \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $B_w \triangleq n\tau_{\text{eff}} \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $E_w \triangleq n \max_i w_i$, $C_w \triangleq \sum_{i=1}^n w_i(\|\boldsymbol{a}_i\|_2^2 - [\alpha_i^{(t,\tau_i-1)}]^2)$, and $D \triangleq \max_i(\|\mathbf{a}_{i,-1}\|_1^2 + \beta_L^2 \|\mathbf{a}_{i,-1}\|_2^2)$, $F_w \triangleq \frac{n(n-P)}{P(n-1)} \sum_{i=1}^n w_i^2$. With the following parameter values:

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

where $\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i$, we can further simplify to

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

Proof. We sum the bound in Lemma C.1 over $t = 0$ to $T-1$ and rearrange the terms to get

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

$$
\begin{aligned}
&+ 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^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] \\
&+ 32\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 + 64L_f^2 M_{\mathbf{a}_{-1}}\left([\eta_x^c]^2 G_{\mathbf{x}}^2 + [\eta_y^c]^2\sigma_G^2\right) \qquad \text{(From Lemma C.2)} \\
&+ 128[\eta_y^c]^2 L_f^3 M_{\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^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] \\
&\qquad \text{(where } \bar{\Delta}_{\widetilde{\Phi}} \triangleq \widetilde{\Phi}_{1/2L_f}(\mathbf{x}_0) - \min_{\mathbf{x}}\widetilde{\Phi}_{1/2L_f}(\mathbf{x})\text{)} \\
&+ 32\left([\eta_x^c]^2 + [\eta_y^c]^2\right) 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(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^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)} \\
&\qquad \text{(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] \\
&\qquad \text{(From Lemma C.3; using } A_m \leq \min\{\tfrac{1}{2}, \tfrac{1}{16\beta_G^2}\}\text{)} \\
&+ 72[\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 + 2\sigma_G^2 M_{\mathbf{a}_{-1}}\right]. \qquad (45)
\end{aligned}
$$

We can simplify the notation using the constants $A_w \triangleq n\tau_{\text{eff}}\sum_{i=1}^{n}\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}$, $B_w \triangleq \tau_{\text{eff}} n\left(\max_i \frac{w_i\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\right)$, $E_w \triangleq n\max_i w_i$, $C_w \triangleq \sum_{i=1}^{n} w_i\left\|\mathbf{a}_{i,-1}\right\|_2^2$, $D \triangleq M_{\mathbf{a}_{-1}}$, $F_w \triangleq \frac{n}{P}\sum_{i=1}^{n} w_i^2$ and drop the numerical constants, for simplicity, to get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \\
&\lesssim \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + \tau_{\text{eff}}\gamma_x^s L_f\left[\frac{A_w}{P\tau_{\text{eff}}}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)} + \frac{(n-P)}{(n-1)}F_w\right)\right] \\
&\quad + \tau_{\text{eff}}\gamma_x^s L_f G_{\mathbf{x}}(S-1)\sqrt{\frac{A_w}{P\tau_{\text{eff}}}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)} + \frac{(n-P)}{(n-1)}F_w\right)} \\
&\quad + \frac{L_f R}{\tau_{\text{eff}}\gamma_y^s S} + \frac{\gamma_y^s L_f}{P}\left(A_w\sigma_L^2 + \sigma_G^2\left(\tau_{\text{eff}}\frac{n-P}{n-1}E_w + B_w\beta_L^2\right)\right) + \left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2\left[C_w\sigma_L^2 + D\left(G_{\mathbf{x}}^2 + \sigma_G^2\right)\right] \\
&= \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + \tau_{\text{eff}}\gamma_x^s L_f\mathcal{I}_1^2 + \frac{\gamma_y^s L_f\mathcal{I}_2}{P} + L_f\left[\tau_{\text{eff}}\gamma_x^s G_{\mathbf{x}}(S-1)\mathcal{I}_1 + \frac{R}{\tau_{\text{eff}}\gamma_y^s S}\right] \\
&\quad + \left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2\left[C_w\sigma_L^2 + D(G_{\mathbf{x}}^2 + \sigma_G^2)\right], \qquad (46)
\end{aligned}
$$

where in (46), to simplify notation, we have defined $\mathcal{I}_1 \triangleq \sqrt{\frac{A_w}{P\tau_{\text{eff}}}\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + G_{\mathbf{x}}^2\left(\frac{n(P-1)}{P(n-1)} + \frac{(n-P)}{(n-1)}F_w\right)}$, $\mathcal{I}_2 \triangleq A_w\sigma_L^2 + (B_w\beta_L^2 + \tau_{\text{eff}}\frac{n-P}{n-1}E_w)\sigma_G^2$.

Next, we optimize the algorithm parameters $S, \gamma_x^s, \gamma_y^s, \eta_y^c, \eta_y^c$ to achieve a tight bound on (46). If $R = 0$, we let $S = 1$. Else, let $S = \sqrt{\frac{R}{\tau_{\text{eff}}^2 \gamma_x^s \gamma_y^s G_{\mathbf{x}} \mathcal{I}_1}}$. Substituting this in (46), we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \lesssim \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + \tau_{\text{eff}}\gamma_x^s L_f \mathcal{I}_1^2 + \frac{\gamma_y^s L_f \mathcal{I}_2}{P} + L_f\sqrt{\frac{R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1}{\gamma_y^s}} + \left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2 \left[C_w\sigma_L^2 + D(G_{\mathbf{x}}^2 + \sigma_G^2)\right], \tag{47}$$

Next, we focus on the terms in (47) containing γ_y^s: $L_f\left[\frac{\gamma_y^s \mathcal{I}_2}{P} + \sqrt{\frac{R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1}{\gamma_y^s}}\right]$. To optimize these, we choose $\gamma_y^s = \left(\frac{P}{2\mathcal{I}_2}\right)^{2/3}(R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1)^{1/3}$. Substituting in (47), we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \lesssim \frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + \tau_{\text{eff}}\gamma_x^s L_f \mathcal{I}_1^2 + L_f\left(\frac{\mathcal{I}_2}{P}R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1\right)^{1/3} + \left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2 \left[C_w\sigma_L^2 + D(G_{\mathbf{x}}^2 + \sigma_G^2)\right], \tag{48}$$

Finally, we focus on the terms in (48) containing γ_x^s: $\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}\gamma_x^s T} + L_f\left(\frac{\mathcal{I}_2}{P}R\gamma_x^s G_{\mathbf{x}}\mathcal{I}_1\right)^{1/3}$. We ignore the higher order linear term. With $\gamma_x^s = \left(\frac{3\bar{\Delta}_{\widetilde{\Phi}}}{\tau_{\text{eff}}L_f T}\right)^{3/4}\left(\frac{\mathcal{I}_2}{P}RG_{\mathbf{x}}\mathcal{I}_1\right)^{-1/4}$, and absorbing numerical constants inside $\mathcal{O}(\cdot)$ we get,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \\
&\leq \mathcal{O}\left(\frac{\left(\bar{\Delta}_{\widetilde{\Phi}}\mathcal{I}_1\mathcal{I}_2\right)^{1/4}}{(\tau_{\text{eff}}PT)^{1/4}}\right) + \mathcal{O}\left(\frac{(\bar{\Delta}_{\widetilde{\Phi}}\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}\frac{\mathcal{I}_1^2}{(\mathcal{I}_1\mathcal{I}_2)^{1/4}}\right) + \mathcal{O}\left(\left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2\left[C_w\sigma_L^2 + D(G_{\mathbf{x}}^2 + \sigma_G^2)\right]\right) \\
&\leq \mathcal{O}\left(\frac{\left(\bar{\Delta}_{\widetilde{\Phi}}\sqrt{G_{\mathbf{x}}^2(1+F_w)}A_w\sigma_L^2\right)^{1/4}}{(\tau_{\text{eff}}PT)^{1/4}}\right) + \mathcal{O}\left(\frac{\left(\bar{\Delta}_{\widetilde{\Phi}}\sqrt{G_{\mathbf{x}}^2(1+F_w)}\tau_{\text{eff}}\frac{n-P}{n-1}E_w\sigma_G^2\right)^{1/4}}{(\tau_{\text{eff}}PT)^{1/4}}\right) \\
&\quad + \mathcal{O}\left(\frac{(\bar{\Delta}_{\widetilde{\Phi}}\tau_{\text{eff}}P)^{1/4}}{T^{3/4}}\frac{G_{\mathbf{x}}^2\left(1+\frac{n-P}{n-1}F_w\right)}{\left(A_w + \tau_{\text{eff}}\frac{n-P}{n-1}E_w\right)^{1/4}}\right) + \mathcal{O}\left(\left([\eta_x^c]^2 + [\eta_y^c]^2\right) L_f^2\left[C_w\sigma_L^2 + D(G_{\mathbf{x}}^2 + \sigma_G^2)\right]\right).
\end{aligned} \tag{49}$$

Lastly, we specify the algorithm parameters in terms of $n, T, \tau_{\text{eff}}, \bar{\tau}$.

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

Finally, choosing the client learning rates $\eta_x^c = \eta_y^c = \frac{1}{L_f\bar{\tau}T^{3/8}}$, we get the result. □

Convergence in terms of F

Proof of Corollary 2.1. Following Lin et al. (2020a), we define

$$\begin{aligned}
\widetilde{\Phi}_{1/2L_f}(\mathbf{x}) &\triangleq \min_{\mathbf{x}'}\left\{\widetilde{\Phi}(\mathbf{x}') + L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\}; \qquad & \widetilde{\mathbf{x}} &\triangleq \arg\min_{\mathbf{x}'}\left\{\widetilde{\Phi}(\mathbf{x}') + L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\}, \\
\Phi_{1/2L_f}(\mathbf{x}) &\triangleq \min_{\mathbf{x}'}\left\{\Phi(\mathbf{x}') + L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\}; \qquad & \bar{\mathbf{x}} &\triangleq \arg\min_{\mathbf{x}'}\left\{\Phi(\mathbf{x}') + L_f\left\|\mathbf{x}'-\mathbf{x}\right\|^2\right\}.
\end{aligned} \tag{50}$$

Also, it follows from Lemma 2.2 in Davis & Drusvyatskiy (2019) that $\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}) = 2L_f(\mathbf{x}-\widetilde{\mathbf{x}})$ and $\nabla\Phi_{1/2L_f}(\mathbf{x}) = 2L_f(\mathbf{x}-\bar{\mathbf{x}})$. Therefore,

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

Consequently, we obtain

$$\begin{aligned}\min_{t\in[T]}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla\Phi_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \\ &\leq \frac{2}{T}\sum_{t=0}^{T-1}\left[\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 + 4L_f^2\left\|\widetilde{\mathbf{x}}^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2\right].\end{aligned} \tag{51}$$

where $\widetilde{\mathbf{x}}^{(t)}, \bar{\mathbf{x}}^{(t)}$ follow the same definition as in (50), with $\mathbf{x}$ replaced with $\mathbf{x}^{(t)}$. □

Proof of Corollary 2.2. If clients are weighted equally ($w_i = p_i = 1/n$ for all i), with each carrying out τ steps of local SGDA+, then (6) reduces to

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

- For full client participation, this reduces to

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

  To reach an ϵ-stationary point, assuming $n\tau \leq T$, the per-client gradient complexity is $T\tau = \mathcal{O}\left(\frac{1}{n\epsilon^8}\right)$. Since $\tau \leq T/n$, the minimum number of communication rounds required is $T = \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$.

- For partial participation, $\mathcal{O}\Big(\Big(\frac{n-P}{n-1}\cdot\frac{1}{PT}\Big)^{1/4}\Big)$ is the dominant term, and we do not get any convergence benefit of multiple local updates. Consequently, per-gradient client complexity and number of communication rounds are both $T\tau = \mathcal{O}\left(\frac{1}{P\epsilon^8}\right)$, for $\tau = \mathcal{O}(1)$. However, if the data across clients comes from identical distributions ($\sigma_G = 0$), then we recover per-client gradient complexity of $\mathcal{O}\left(\frac{1}{P\epsilon^8}\right)$, and number of communication rounds $= \mathcal{O}\left(\frac{1}{\epsilon^4}\right)$.

□

Special Cases

- Centralized, deterministic case ($\sigma_L = \sigma_G = 0, \beta_G = 1, \tau_{\text{eff}} = n = 1$): in this case $A_w = B_w = 1, C_w = D = 0$. Also, $\mathcal{I}_1 = G_{\mathbf{x}}\sqrt{\beta_L^2+1}, \mathcal{I}_2 = 0$. The bound in (45) reduces to

  $$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \mathcal{O}\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\gamma_x^s T} + \gamma_x^s L_f G_{\mathbf{x}}^2\left[(\beta_L^2+1) + (S-1)\sqrt{\beta_L^2+1}\right] + \frac{L_f R}{\gamma_y^s S}\right). \tag{52}$$

  For $\beta_L = 0$, (52) yields the convergence result in Lin et al. (2020a).

- Single node, stochastic case ($\sigma_G = 0, \beta_G = 1, \tau_{\text{eff}} = n = 1$): in this case $A_w = B_w = 1, C_w = D = 0$. Also, $\mathcal{I}_1 = \sqrt{\sigma_L^2 + (\beta_L^2+1)G_{\mathbf{x}}^2}, \mathcal{I}_2 = \sigma_L^2$. The bound in (45) reduces to

  $$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \Theta\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\gamma_x^s T} + \gamma_x^s L_f(\sigma_L^2 + (\beta_L^2+1)G_{\mathbf{x}}^2) + \gamma_y^s L_f\sigma_L^2\right)$$

$$+\Theta\left(L_f\left[\gamma_x^s G_{\mathbf{x}}(S-1)\sqrt{\sigma_L^2+(\beta_L^2+1)G_{\mathbf{x}}^2}+\frac{R}{\gamma_y^s S}\right]\right). \tag{53}$$

Again, for $\beta_L = 0$, (53) yields the convergence result in Lin et al. (2020a).

- Multiple equally weighted ($w_i = 1/n, \forall\, i \in [n]$) clients, full client participation, stochastic case with synchronous client updates ($\tau_{\text{eff}} = 1$): in this case $A_w = 1, B_w = 1, C_w = D = 0, 0$. The bound in (45) reduces to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \Theta\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\gamma_x^s T}+\gamma_x^s L_f\left(G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2}{n}\right)\right)$$
$$+\Theta\left(\frac{\gamma_y^s L_f}{n}(\sigma_L^2+\sigma_G^2\beta_L^2)+\gamma_x^s L_f G_{\mathbf{x}}(S-1)\sqrt{G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2}{n}}+\frac{RL_f}{\gamma_y^s S}\right), \tag{54}$$

Note that unlike existing analyses of synchronous update algorithms Woodworth et al. (2020); Yun et al. (2022); Sharma et al. (2022), the bound in (54) depends on the inter-client heterogeneity σ_G^2. This is due to the more general noise assumption (Assumption 3). In the existing works, β_L^2 is assumed zero, in which case, the bound in (54) is also independent of σ_G^2. See Appendix A.2 for a more detailed explanation.

- Multiple, equally weighted ($w_i = 1/n, \forall\, i \in [n]$) clients, full client participation, multiple, but equal number of client updates ($\tau_i = \tau_{\text{eff}} = \tau, \forall\, i \in [n]$). In this case $A_w = B_w = 1, C_w = \tau - 1, D = (\tau-1)(\tau-1+\beta_L^2)$. The bound in (45) then reduces to

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t)})\right\|^2 \leq \Theta\left(\frac{\bar{\Delta}_{\widetilde{\Phi}}}{\tau\gamma_x^s T}+\tau\gamma_x^s L_f\left(G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2}{n\tau}\right)+\frac{\gamma_y^s L_f(\sigma_L^2+\beta_L^2\sigma_G^2)}{n}\right) \tag{55}$$
$$+\Theta\left(L_f\left[\tau\gamma_x^s G_{\mathbf{x}}(S-1)\sqrt{G_{\mathbf{x}}^2+\frac{\sigma_L^2+\beta_L^2 G_{\mathbf{x}}^2}{n\tau}}+\frac{R}{\tau\gamma_y^s S}\right]+(\tau-1)\left([\eta_x^c]^2+[\eta_y^c]^2\right)L_f^2\left[\sigma_L^2+(\tau-1+\beta_L^2)(G_{\mathbf{x}}^2+\sigma_G^2)\right]\right).$$

For $\beta_L = \beta_G = 0$, this setting reduces to the one considered in Sharma et al. (2022). However, as stated earlier, our bound on the local update error is tighter.

C.3 Proofs of the Intermediate Lemmas

Proof of Lemma C.1. Using the definition in (50) $\bar{\mathbf{x}}^{(t)} = \arg\min_{\mathbf{x}} \widetilde{\Phi}(\mathbf{x}) + L_f\left\|\mathbf{x}-\mathbf{x}^{(t)}\right\|^2$. Also, note that

$$\widetilde{\Phi}_{1/2L_f}(\mathbf{x}^{(t+1)}) \leq \widetilde{\Phi}(\bar{\mathbf{x}}^{(t)}) + L_f\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t+1)}\right\|^2. \tag{56}$$

Using the $\mathbf{x}$ updates in (43),

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

$$\leq \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2+\tau_{\text{eff}}^2[\gamma_x^s]^2\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t)}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t)}\right\|^2 \tag{57}$$

$$+\tau_{\text{eff}}\gamma_x^s\mathbb{E}\left[\frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2+2L_f\sum_{i=1}^n\frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1}a_i^{(k)}(\tau_i)\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)+2\left\langle\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)},\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\rangle\right], \tag{58}$$

where (58) follows from L_f-smoothness (Assumption 1) and Jensen's inequality. From (18), (19), we can bound $\mathbb{E}\|\sum_{i\in\mathcal{C}^{(t')}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t')}\|^2$ as follows.

$$\begin{aligned}\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t')}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t')}\right\|^2 &\leq \frac{n}{P}\sum_{i=1}^n\frac{w_i^2}{\|\boldsymbol{a}_i\|_1^2}\sum_{k=0}^{\tau_i-1}[a_i^{(k)}(\tau_i)]^2\left[\sigma_L^2+\beta_L^2\mathbb{E}\left\|\nabla_x f_i(\mathbf{x}_i^{(t,k)},\mathbf{y}_i^{(t,k)})\right\|^2\right]\\ &\quad+\frac{n}{P}\left(\frac{P-1}{n-1}\right)\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2+\frac{n}{P}\frac{n-P}{n-1}\sum_{i=1}^n w_i^2\mathbb{E}\left\|\mathbf{h}_{\mathbf{x},i}^{(t)}\right\|^2\\ &\leq \frac{n}{P}\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}\left(\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2\right)+\frac{n(P-1)}{P(n-1)}G_{\mathbf{x}}^2+\frac{n(n-P)}{P(n-1)}G_{\mathbf{x}}^2\sum_{i=1}^n w_i^2,\end{aligned} \tag{59}$$

where the final inequality by using Assumption 8. Next, we bound the inner product term in (58). Using L_f-smoothness of F (Assumption 1):

$$\begin{aligned}&\mathbb{E}\left\langle\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)},\nabla_x\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\rangle\leq\mathbb{E}\left[\widetilde{F}(\bar{\mathbf{x}}^{(t)},\mathbf{y}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})+\frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2\right]\\ &\leq\mathbb{E}\left[\widetilde{\Phi}(\bar{\mathbf{x}}^{(t)})+L_f\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2\right]-\mathbb{E}\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\frac{L_f}{2}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2\\ &\leq\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})+L_f\left\|\mathbf{x}^{(t)}-\mathbf{x}^{(t)}\right\|^2\right]-\mathbb{E}\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\frac{L_f}{2}\mathbb{E}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2 \qquad \text{(by definition of } \bar{\mathbf{x}}^{(t)})\\ &\leq\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\frac{L_f}{2}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\right\|^2\right].\end{aligned} \tag{60}$$

Substituting the bounds from (58) and (60) into (56), we get

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

where we use $\nabla\widetilde{\Phi}_{1/2L_f}(\mathbf{x})=2L_f(\mathbf{x}-\bar{\mathbf{x}})$ from (50). □

Proof of Lemma C.2. We use the client update equations for individual iterates in (42). To bound $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$, first we bound the $\mathbf{x}$-error $\mathbb{E}\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2$. Starting from (24), using Assumption 8, for $1 \le k \le \tau_i$,

$$\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\mathbf{x}_i^{(t,k)} - \mathbf{x}^{(t)}\right\|^2 \le \frac{[\eta_x^c]^2}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\left[\sum_{j=0}^{k-1}[a_i^{(j)}(k)]^2\left(\sigma_L^2 + \beta_L^2 G_{\mathbf{x}}^2\right) + \left(\sum_{j=0}^{k-1} a_i^{(j)}(k)\right)\sum_{j=0}^{k-1} a_i^{(j)}(k) G_{\mathbf{x}}^2\right]$$
$$\le [\eta_x^c]^2\left[\sigma_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2 + G_{\mathbf{x}}^2\left(\left\|\mathbf{a}_{i,-1}\right\|_1^2 + \beta_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2\right)\right], \tag{61}$$

where we use (25). Next, we bound $\mathbb{E}\left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2$, using the bound from (29), to get

$$\frac{1}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\mathbb{E}\left\|\mathbf{y}_i^{(t,k)} - \mathbf{y}^{(t)}\right\|^2 \le [\eta_y^c]^2\sigma_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2 + 2[\eta_y^c]^2 L_f^2\left(\left\|\mathbf{a}_{i,-1}\right\|_1 + \beta_L^2\alpha\right)\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{y}}^{(t,k)}(i)$$
$$+ 2[\eta_y^c]^2\left(\left\|\mathbf{a}_{i,-1}\right\|_1^2 + \beta_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2\right)\mathbb{E}\left\|\nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right\|^2. \tag{62}$$

Compared to (29), the difference is the presence of $\Delta_{\mathbf{y}}^{(t,k)}(i)$ in (62), rather than $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i)$. Taking a weighted sum over agents in (62), we get

$$L_f^2\sum_{i=1}^n \frac{w_i}{\|\boldsymbol{a}_i\|_1}\sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i)\Delta_{\mathbf{y}}^{(t,k)}(i) \le 2[\eta_y^c]^2 L_f^2\left[\sigma_L^2\sum_{i=1}^n w_i\left\|\mathbf{a}_{i,-1}\right\|_2^2 + 2M_{\mathbf{a}_{-1}}\left(\beta_G^2\mathbb{E}\left\|\nabla_y\widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right\|^2 + \sigma_G^2\right)\right]. \tag{63}$$

where, we choose η_y^c such that $A_m \triangleq 2L_f^2[\eta_y^c]^2\max_i\|\boldsymbol{a}_i\|_1\left(\|\mathbf{a}_{i,-1}\|_1 + \beta_L^2\alpha\right) \le \frac{1}{2}$, and define $M_{\mathbf{a}_{-1}} \triangleq \max_i\left(\left\|\mathbf{a}_{i,-1}\right\|_1^2 + \beta_L^2\left\|\mathbf{a}_{i,-1}\right\|_2^2\right)$. Next, it follows from L_f-smoothness (Assumption 1) and Lemma A.7 that

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

Subsequently, combining (61) and (63), we get

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

which finishes the proof. □

Proof of Lemma C.3. We define $\mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}) \in \arg\max_{\mathbf{y}}\widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y})$. Then,

$$\mathbb{E}\left\|\mathbf{y}^{(t+1)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 \overset{(43)}{=} \mathbb{E}\left\|\mathbf{y}^{(t)} + \tau_{\text{eff}}\gamma_y^s\mathbf{g}_{\mathbf{y}}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2$$
$$= \mathbb{E}\left\|\mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})\right\|^2 + \tau_{\text{eff}}^2[\gamma_y^s]^2\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2 + 2\tau_{\text{eff}}\gamma_y^s\mathbb{E}\left\langle\mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\rangle. \tag{64}$$

$\mathbb{E}\left\|\mathbf{g}_{\mathbf{y}}^{(t)}\right\|^2$ is bounded in (38). We only need to further bound $\mathbb{E}\left\|\sum_{i=1}^n w_i\mathbf{h}_{\mathbf{y},i}^{(t)}\right\|^2$.

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

$$\leq 2L_f^2 \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \Delta_{\mathbf{y}}^{(t,k)}(i) + 4L_f \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right]. \tag{65}$$

Next, we bound the third term in (64).

$$\begin{aligned}
&\mathbb{E}\left\langle \mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \sum_{i=1}^{n} w_i \mathbf{h}_{\mathbf{y},i}^{(t)} \right\rangle = \mathbb{E}\left\langle \mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) \right\rangle \\
&= \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \mathbb{E}\left[\left\langle \mathbf{y}^{(t)} - \mathbf{y}_i^{(t,k)}, \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) \right\rangle + \left\langle \mathbf{y}_i^{(t,k)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}), \nabla_y f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) \right\rangle\right] \\
&\leq \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \mathbb{E}\Bigg[f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}) - f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) + \frac{L_f}{2} \left\| \mathbf{y}^{(t)} - \mathbf{y}_i^{(t,k)} \right\|^2 && (L_f\text{-smoothness}) \\
&\qquad + f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}_i^{(t,k)}) - f_i(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})) \Bigg] && (\text{Concavity in } \mathbf{y}) \\
&= \frac{L_f}{2} \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \Delta_{\mathbf{y}}^{(t,k)}(i) - \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right].
\end{aligned} \tag{66}$$

Substituting (38), (65), (66) in (64), we get

$$\begin{aligned}
&\mathbb{E}\left\| \mathbf{y}^{(t+1)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}) \right\|^2 \\
&\leq \mathbb{E}\left\| \mathbf{y}^{(t)} - \mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}) \right\|^2 + \tau_{\text{eff}}^2 [\gamma_y^s]^2 \left[\frac{\sigma_L^2 n}{P} \sum_{i=1}^{n} \frac{w_i^2 \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} + \frac{2\sigma_G^2 n}{P} \left(\frac{n-P}{n-1} \max_i w_i + \beta_L^2 \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} \right) \right] \\
&\quad - 2\tau_{\text{eff}} \gamma_y^s \left(1 - 2\tau_{\text{eff}} \gamma_y^s L_f \frac{n}{P} \left[\frac{P-1}{n-1} + \beta_G^2 \left(\frac{n-P}{n-1} \max_i w_i + \beta_L^2 \max_i \frac{w_i \|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2} \right) \right] \right) \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)}) \right] \\
&\quad + \tau_{\text{eff}} \gamma_y^s L_f \left[1 + 2\tau_{\text{eff}} \gamma_y^s L_f \frac{n}{P} \left(\frac{P-1}{n-1} + \frac{n-P}{n-1} \max_i w_i + \beta_L^2 \max_{i,k} \frac{w_i a_i^{(k)}(\tau_i)}{\|\boldsymbol{a}_i\|_1} \right) \right] \sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} [a_i^{(k)}(\tau_i)] \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i),
\end{aligned} \tag{67}$$

since $\Delta_{\mathbf{y}}^{(t,k)} \leq \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}$. Using the bound on $\Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}$ from Lemma C.2,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{w_i}{\|\boldsymbol{a}_i\|_1} \sum_{k=0}^{\tau_i-1} a_i^{(k)}(\tau_i) \Delta_{\mathbf{x},\mathbf{y}}^{(t,k)}(i) &\leq 2\left([\eta_x^c]^2 + [\eta_y^c]^2\right) \sigma_L^2 \sum_{i=1}^{n} w_i \|\mathbf{a}_{i,-1}\|_2^2 + 4M_{\mathbf{a}_{-1}} \left([\eta_x^c]^2 G_{\mathbf{x}}^2 + [\eta_y^c]^2 \sigma_G^2\right) \\
&\quad + 8[\eta_y^c]^2 L_f M_{\mathbf{a}_{-1}} \beta_G^2 \mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)}) - \widetilde{F}(\widehat{\mathbf{x}}^{(s)}, \mathbf{y}^{(t)})\right].
\end{aligned} \tag{68}$$

We substitute (68) in (67), and simplify the terms using the choice of γ_y^s, η_y^c to get

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

using γ_y^s, η_y^c that satisfy

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

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

Then the coefficient of $\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)})-\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^{(t)})\right]$ can we bounded by $-\tau_{\text{eff}}\gamma_y^s$. Consequently, by rearranging the terms and summing over t, we get the result.

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

□

Proof of Lemma C.4. Let $t=sS, sS+1,\ldots,(s+1)S-1$, where k is a positive integer. Let $\widehat{\mathbf{x}}^{(s)}$ is the latest snapshot iterate for the $\mathbf{y}$-update in Algorithm 1-Fed-Norm-SGDA+ . Then

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\Phi}(\mathbf{x}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&=\mathbb{E}\left[\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^*(\mathbf{x}^{(t)}))-\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}))+\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}))-\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^{(t)})+\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^{(t)})-\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right]\\
&\leq\mathbb{E}\left[\widetilde{F}(\mathbf{x}^{(t)},\mathbf{y}^*(\mathbf{x}^{(t)}))-\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^*(\mathbf{x}^{(t)}))\right]+\mathbb{E}\left[\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^*(\widehat{\mathbf{x}}^{(s)}))-\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^{(t)})\right]+G_{\mathbf{x}}\mathbb{E}\left\|\mathbf{x}^{(t)}-\widehat{\mathbf{x}}^{(s)}\right\|\\
&\leq 2G_{\mathbf{x}}\mathbb{E}\left\|\mathbf{x}^{(t)}-\widehat{\mathbf{x}}^{(s)}\right\|+\mathbb{E}\left[\widetilde{\Phi}(\widehat{\mathbf{x}}^{(s)})-\widetilde{F}(\widehat{\mathbf{x}}^{(s)},\mathbf{y}^{(t)})\right].
\end{aligned}\tag{69}$$

where, $\mathbf{y}^*(\cdot)\in\arg\max_{\mathbf{y}}\widetilde{F}(\cdot,\mathbf{y})$ and (69) follows from $G_{\mathbf{x}}$-Lipschitz continuity of $F(\cdot,\mathbf{y})$ (Assumption 8). Next, we see that

$$\begin{aligned}
\mathbb{E}\left\|\widehat{\mathbf{x}}^{(s)}-\mathbf{x}^{(t)}\right\| &\leq\sqrt{\mathbb{E}\left\|\widehat{\mathbf{x}}^{(s)}-\mathbf{x}^{(t)}\right\|^2} && \text{(Jensen's inequality)}\\
&\overset{(43)}{=}\sqrt{\mathbb{E}\left\|\tau_{\text{eff}}\gamma_x^s\sum_{t'=sS}^{t-1}\sum_{i\in\mathcal{C}^{(t')}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t')}\right\|^2}\\
&\leq\tau_{\text{eff}}\gamma_x^s\sqrt{(S-1)\sum_{t'=sS}^{t-1}\mathbb{E}\left\|\sum_{i\in\mathcal{C}^{(t')}}\tilde{w}_i\mathbf{g}_{\mathbf{x},i}^{(t')}\right\|^2}\\
&\leq\tau_{\text{eff}}\gamma_x^s(S-1)\sqrt{\frac{n}{P}}\sqrt{\sum_{i=1}^n\frac{w_i^2\|\boldsymbol{a}_i\|_2^2}{\|\boldsymbol{a}_i\|_1^2}(\sigma_L^2+\beta_L^2G_{\mathbf{x}}^2)+G_{\mathbf{x}}^2\left(\frac{P-1}{n-1}+\frac{n-P}{n-1}\sum_{i=1}^n w_i^2\right)}. && \text{(from (59))}
\end{aligned}$$

Using this bound in (69), and summing over t, we get

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

Finally, summing over $s=0$ to $T/S-1$ we get the result. □

C.4 Extending the result for Nonconvex One-Point-Concave (NC-1PC) Functions

Carefully revisting the proof of Theorem 2, we notice that Lemma C.1 and Lemma C.2 do not rely on the concavity assumption. Lemma C.3 does use concavity of local functions $\{f_i\}$. However, it is only needed to derive (66). Further, this only requires concavity of local functions at a global point $\mathbf{y}^*(\widehat{\mathbf{x}}^{(s)})$. Therefore, as mentioned earlier in Remark 10, it holds even for NC-1PC functions. This is an independent result in itself, since we have extended the existing convergence result of local stochastic gradient method for convex minimization (concave maximization) problems, to a much more general one-point-convex minimization (or one-point-convex maximization) problem. Therefore, we restate it here for the more general case.

Lemma C.5 (Local SG updates for One-Point-Concave Maximization)**.** *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 3, 4, 8. Suppose for all $\mathbf{x}$, all the f_i's satisfy Assumption 9 at a common global minimizer $\mathbf{y}^*(\mathbf{x})$, and that $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ for all t. If we run Fed-Norm-SGDA+ with the same conditions on the client and server step-sizes η_y^c, γ_y^s respectively, as in Lemma C.3, then the iterates generated by Fed-Norm-SGDA+ also satisfy the bound in Lemma C.3.*

Next, Lemma C.4 also holds irrespective of concavity. Therefore, the resulting convergence result in Theorem 2 for nonconvex-concave minimax problems holds for a much larger class of functions. We restate the modified theorem statement briefly.

Theorem. *Suppose the local loss functions $\{f_i\}$ satisfy Assumptions 1, 3, 4, 8. Suppose for all $\mathbf{x}$, all the f_i's satisfy Assumption 9 at a common global minimizer $\mathbf{y}^*(\mathbf{x})$, and that $\left\|\mathbf{y}^{(t)}\right\|^2 \leq R$ for all t. If we run Algorithm 1-Fed-Norm-SGDA+ with the same conditions on the client and server step-sizes η_y^c, γ_y^s respectively, as in Theorem 3, then the iterates generated by Algorithm 1-Fed-Norm-SGDA+ also satisfy the bound in Theorem 3.*

Remark 11. Again, choosing client weights $\{w_i\}$ the same as in the original global objective $\{p_i\}$, we get convergence in terms of the original objective F.

D Additional Experiments

For communicating parameters and related information amongst the clients, ethernet connections were used. Our algorithm was implemented using parallel training tools in PyTorch 1.0.0 and Python 3.6.3.

For both robust NN Training and fair classification experiments, we use batch-size of 32 in all the algorithms. Momentum parameter 0.9 is used only in Momentum Local SGDA(+).

Robust NN Training. Here we further explore performance of Fed-Norm-SGDA+ on the robust NN training problem. We use VGG-11 model to classify CIFAR10 dataset. In Figure 6, we show the training loss curves corresponding to the results in Figure 3 on varying number of local steps. Similarly, in Figure 6, we show the training loss curves corresponding to Figure 4 on the effect of partial participation. In Figure 7, we demonstrate the effect of increasing data heterogeneity across clients, whle in Figure 9 we show the advantage of using multiple clients for the federated minimax problem. With k-fold increase in n, we observe an almost k-fold drop in the number of communication rounds needed to reach a target test accuracy (70% here.).

We use batch-size of 32. Momentum parameter 0.9 is used only in Local SGDA+(M).

Table 3: Parameter values for experiments in robust NN training experiments.

Communication rounds	1-100	101-200	>200
Client Learning Rate (η_y^c)	0.02	2×10^{-3}	2×10^{-4}
Client Learning Rate (η_x^c)	0.016	1.6×10^{-3}	1.6×10^{-4}
Server Learning Rate ($\gamma_x^s = \gamma_y^s$)	1	1	1

Fair Classification We also demonstrate the impact of partial client participation in the fair classification problem. Figure 10 complements Figure 10 in the main paper, evaluating fairness of a VGG11 model on

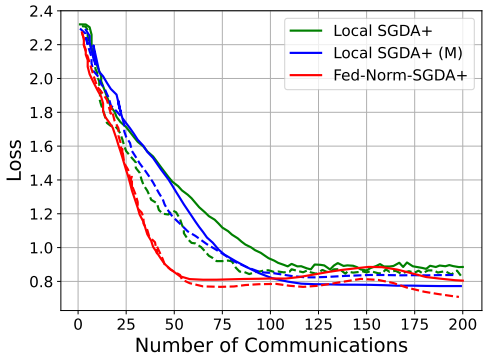

Figure 6: Comparison of the effect of heterogeneous number of local updates $\{\tau_i\}$ on the performance of Fed-Norm-SGDA+ (Algorithm 1), Local SGDA+, and Local SGDA+ with momentum, while solving (7) on CIFAR10 dataset, with VGG11 model. The solid (dashed) curves are for $E = 5$ ($E = 7$), and $\alpha = 0.1$.

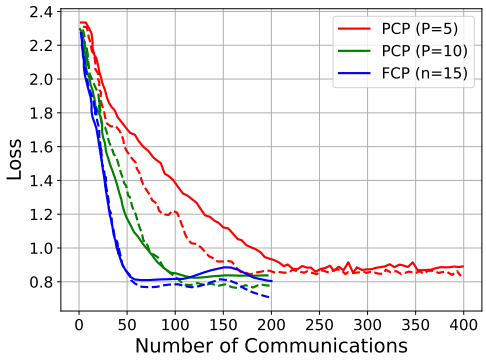

Figure 7: Comparison of the effects of partial client participation (PCP) on the performance of Fed-Norm-SGDA+, for the robust NN training problem on the CIFAR10 dataset, with the VGG11 model. The figure shows the robust test accuracy. The solid (dashed) curves are for $\alpha = 0.1$ ($\alpha = 1.0$).

CIFAR10 dataset. We have plotted the test accuracy of the model over the worst distribution. With an increasing number of participating clients, the performance consistently improves.

Batch-size of 32 is used. Momentum parameter 0.9 is used only in Local SGDA (M).

Table 4: Parameter values for experiments in fair classification experiments.

Client Learning Rate (η_y^c)	0.02
Client Learning Rate (η_x^c)	0.016
Server Learning Rate ($\gamma_x^s = \gamma_y^s$)	1

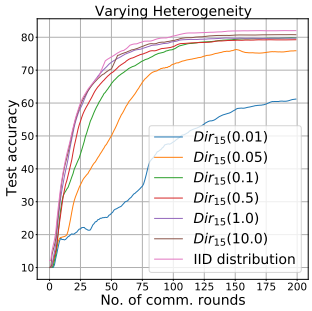

Figure 8: Effect of inter-client data heterogeneity (quantified by α) on the performance of Fed-Norm-SGDA+ in a robust NN training task.

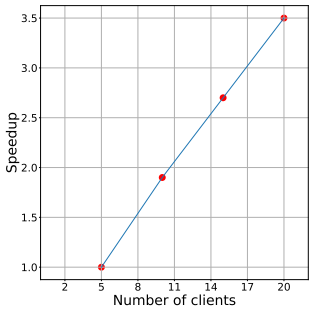

Figure 9: Effect of increasing client-set on the performance of Fed-Norm-SGDA+ in a robust NN training task.

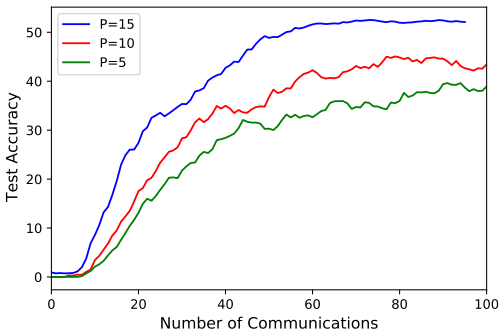

Figure 10: Effect of partial client participation on the performance of Fed-Norm-SGDA in a fair image classification task.