Stochastic Smoothed Gradient Descent Ascent for Federated Minimax Optimization

Wei Shen University of Virginia
Minhui Huang Meta AI
Jiawei Zhang MIT
Cong Shen University of Virginia

Abstract

In recent years, federated minimax optimization has attracted growing interest due to its extensive applications in various machine learning tasks. While Smoothed Alternative Gradient Descent Ascent (Smoothed-AGDA) has proved successful in centralized nonconvex minimax optimization, how and whether smoothing techniques could be helpful in a federated setting remains unexplored. In this paper, we propose a new algorithm termed Federated Stochastic Smoothed Gradient Descent Ascent (`FESS-GDA`), which utilizes the smoothing technique for federated minimax optimization. We prove that `FESS-GDA` can be uniformly applied to solve several classes of federated minimax problems and prove new or better analytical convergence results for these settings. We showcase the practical efficiency of `FESS-GDA` in practical federated learning tasks of training generative adversarial networks (GANs) and fair classification.

1 INTRODUCTION

Minimax optimization is widely encountered in modern machine learning tasks such as generative adversarial networks (GANs) (Goodfellow et al., 2014a), AUC maximization (Liu et al., 2019), reinforcement learning (Zhang et al., 2021), adversarial training (Goodfellow et al., 2014b), and fair machine learning (Nouiehed et al., 2019). In recent years, many progresses on minimax optimization problems have been reported, with the majority focusing on solutions at a single client level. However, modern machine learning tasks usually demand a huge amount of data. A significant portion of this data may be sensitive, rendering it unsuitable for sharing with servers due to privacy concerns (Léauté and Faltings, 2013). Furthermore, data sourced from edge devices can be hindered by the limited communication capabilities with the server. To preserve data privacy and to address communication issues, federated learning (FL) was proposed (McMahan et al., 2017). In FL, clients do not send their data directly to the server. Instead, each client trains its model locally using its own data. Periodically, clients communicate with the server, sending their models for aggregation. The server then returns the updated model to the clients.

Solutions and analyses for federated minimax problems have been developed in recent years. Some focus on convex-concave problems (Deng et al., 2020a; Hou et al., 2021; Sun and Wei, 2022), and others are devoted to more general nonconvex minimax problems (Deng and Mahdavi, 2021; Sharma et al., 2022, 2023). Because the objective functions are usually nonconvex in the min variables for many practical applications, we mainly focus on federated nonconvex minimax problems in this paper.

Gradient descent ascent (GDA) and its stochastic version stochastic gradient descent ascent (SGDA) are the simplest single-loop algorithms for centralized minimax problems. Most existing federated minimax algorithms are extensions of GDA (SGDA) to the federated setting, i.e. Local SGDA (Deng and Mahdavi, 2021), Fed-Norm-SGDA (Sharma et al., 2022). Zhang et al. (2020) propose Smoothed-AGDA, a single-loop algorithm utilizing the smoothing technique, and prove that it has a faster convergence rate for centralized nonconvex-concave problems compared with GDA. Yang et al. (2022b) then prove that Smoothed-AGDA and its stochastic version Stochastic Smoothed-AGDA also have faster convergence rates for centralized nonconvex-PL (Polyak-Lojasiewicz) problems compared with GDA (SGDA). A natural question arises: **Can we utilize the smoothing techniques to design a faster algorithm for federated nonconvex**

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Table 1: Comparison of per-client sample complexity and communication complexity for different classes of nonconvex minimax problems. For comparison, we only give the convergence results for finding an ϵ-stationary point of Φ (Definition 2.2) for NC-PL and of $\Phi_{1/2l}$ (Definition 2.4) for NC-1PC under full client participation ($m = M$). We also provide convergence results of finding an ϵ-stationary point of f (Definition 2.1), and consider partial client participation ($m < M$) in our paper. $\kappa := l/\mu$ is the conditional number.

Algorithms	Partial Client Participation	Full Client Participation (FCP)	
		Per-client Sample complexity	Communication complexity
Nonconvex-Strongly-Concave (NC-SC)/ Nonconvex-PL (NC-PL)			
Local SGDA (Sharma et al., 2022)	✗	$O(\kappa^4 m^{-1}\epsilon^{-4})$	$O(\kappa^3\epsilon^{-3})$
SAGDA (Yang et al., 2022a)	✓	$O(\kappa^4 m^{-1}\epsilon^{-4})$	$O(\kappa^2\epsilon^{-2})$
Fed-Norm-SGDA (Sharma et al., 2023)	✓	$O(\kappa^4 m^{-1}\epsilon^{-4})$	$O(\kappa^2\epsilon^{-2})$
FedSGDA-M[a] (Wu et al., 2023)	✗	$O(\kappa^3 m^{-1}\epsilon^{-3})$	$O(\kappa^2\epsilon^{-2})$
FESS-GDA Corollary 3.1	✓	$O(\kappa^2 m^{-1}\epsilon^{-4})$	$O(\kappa\epsilon^{-2})$
Nonconvex-One-Point-Concave (NC-1PC)			
Local SGDA+[b] (Sharma et al., 2022)	✗	$O(\epsilon^{-8})$	$O(\epsilon^{-7})$
Fed-Norm-SGDA+[b c] (Sharma et al., 2023)	✓	$O(m^{-1}\epsilon^{-8})$	$O(\epsilon^{-4})$
FESS-GDA Theorem 3.2	✓	$O(m^{-1}\epsilon^{-8})$	$O(\epsilon^{-4})$
Nonconvex-Concave (NC-C)			
Local SGDA+[b] (Sharma et al., 2022)	✗	$O(m^{-1}\epsilon^{-8})$	$O(\epsilon^{-7})$
Fed-Norm-SGDA+[b] (Sharma et al., 2023)	✓	$O(m^{-1}\epsilon^{-8})$	$O(\epsilon^{-4})$
FedSGDA+ (Wu et al., 2023)	✗	$O(m^{-1}\epsilon^{-8})$	$O(\epsilon^{-6})$
FESS-GDA[d] Theorem 3.2	✓	$O(m^{-1}\epsilon^{-8})$	$O(\epsilon^{-4})$
Objective function has a form of (2) (A special case of NC-C problems)			
FESS-GDA Theorem 3.4	✓	$O(m^{-1}\epsilon^{-4})$	$O(\epsilon^{-2})$

[a] Their better performance comes from using additional variance reduction, while we do not.
[b] Their proofs need additional assumptions that each local loss function f_i also satisfies the NC-C (NC-1PC) condition, while ours only needs the global loss function f to be NC-C (NC-1PC). They also assume $\|y_t\|^2 \le D$, but do not mention how to guarantee this. We use a projection operator in our algorithm to guarantee this.
[c] Their proof requires additional assumption that each local loss function f_i satisfies the one-point-concave condition with a common global minimizer $y^*(x)$.
[d] We have better convergence results for the NC-C setting of finding an ϵ-stationary point of f; see Theorem 3.3 for details.

minimax optimization?

Furthermore, in the current literature, usually two different algorithms (such as Local SGDA and Local SGDA+ (Deng and Mahdavi, 2021; Sharma et al., 2022)) are needed for different nonconvex minimax settings, which limits their practical applicability. Another question thus arises: **Can we design a single, uniformly applicable algorithm for federated nonconvex minimax optimization?**

1.1 Problem Setting

In this paper, we study the federated minimax optimization problems in the following form:

$$\min_{x\in X}\max_{y\in Y}\left\{f(x,y)=\frac{1}{M}\sum_{i=1}^{M}f_i(x,y)\right\}, \tag{1}$$

where $X=\mathbb{R}^{d_1}, Y\subseteq\mathbb{R}^{d_2}$, M is the number of clients, $f_i(x,y)=\mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f(x,y;\xi_i)]$ is the local loss function at client i, and $f(x,y;\xi_i)$ denotes the loss for the data point ξ_i, sampled from the local data distribution $\mathcal{D}_i$ at client i.

For the nonconvex-concave setting, we also consider a special case:

$$\min_{x\in X}\max_{y\in Y} f(x,y)=\min_{x\in X}\max_{y\in Y} F(x)^T y, \tag{2}$$

where $Y=\{(y_1,...,y_n)^T|\sum_{i=1}^n, y_i=1, y_i\ge 0\}$ and $F(x)=(f_1(x),...,f_n(x))^T$ is a mapping from $X=\mathbb{R}^{d_1}$ to $\mathbb{R}^n$. Note that (2) is equivalent to the problem of minimizing the point-wise maximum of a finite collection of functions:

$$\min_{x}\max_{1\le i\le n} f_i(x). \tag{3}$$

Problems in the form of (2) and (3) commonly appear in practical applications such as adversarial training Nouiehed et al. (2019); Madry et al. (2017) and fairness training Nouiehed et al. (2019).

Table 2: Comparison of per-client sample complexity and communication complexity needed to find (x_T, y_T) that satisfy $\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq \epsilon^2$. We use $\tilde{O}$ to hide logarithmic terms.

Algorithms	Type	Partial Client Participation	Data Heterogeneity	Per-client Sample complexity	Communication complexity
Local SGDA[a] Deng and Mahdavi (2021)	SC-SC	✗	✗	$\tilde{O}(M^{-1}\epsilon^{-2})$	$\tilde{O}(M)$
		✗	✓	$O(M^{-1}\epsilon^{-2})$	$O(\epsilon^{-1})$
FESS-GDA Theorem 3.5	PL-PL	✗	✓	$\tilde{O}(M^{-1}\epsilon^{-2})$	$\tilde{O}(1)$
		✓	✗	$\tilde{O}(m^{-1}\epsilon^{-2})$	$\tilde{O}(1)$
		✓	✓	$\tilde{O}(m^{-1}\epsilon^{-2})$	$\tilde{O}(m^{-1}\epsilon^{-2})$

[a] Their proofs need an assumption that each local loss function f_i satisfies the SC-SC condition, while ours only needs the global loss function f to satisfy the PL-PL condition (Assumption 3.7).

1.2 Contributions

We propose a new algorithm termed FEderated Stochastic Smoothed Gradient Descent Ascent (FESS-GDA). We prove that FESS-GDA can be uniformly used to solve several classes of federated nonconvex minimax problems, and prove new or better convergence results for these settings. We summarize our main theoretical results in Tables 1, 2 with the following abbreviations:

SC-SC: Strongly-Convex in x, Strongly-Concave in y,

PL-PL: PL condition in x, PL condition in y (Assumption 3.7),

NC-SC: Nonconvex in x, Strongly-Concave in y,

NC-PL: Nonconvex in x, PL condition in y (Assumption 3.1),

NC-C: Nonconvex in x, Concave in y (Assumption 3.5),

NC-1PC: Nonconvex in x, One-Point-Concave in y (Assumption 3.4).

More specifically, our contributions are the following.

- For NC-PL and NC-SC problems, we prove that FESS-GDA achieves a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\kappa\epsilon^{-2})$ in terms of the stationarity of both f and Φ. The previously best-known results without variance reduction in the federated setting are $O(\kappa^4 m^{-1}\epsilon^{-4})$ per-client sample complexity and $O(\kappa^2\epsilon^{-2})$ communication complexity. We improve these results by a factor of $O(\kappa^2)$ in the sample complexity and a factor of $O(\kappa)$ in the communication complexity.
- To the best of our knowledge, we are the first to prove convergence results of solving (2) under a federated setting. We prove that FESS-GDA has a sample complexity of $O(m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\epsilon^{-2})$ in terms of the stationarity of both f and Φ, which is much better than the complexity we can achieve for general NC-C problems.
- For general NC-C and NC-1PC problems, we prove that FESS-GDA achieves comparable performances as the current state-of-the-art algorithm, but with weaker assumptions. Moreover, we provide additional convergence results for these two settings in terms of the stationarity of f. For the NC-C problems, we prove a best-known per-client sample complexity of $O(m^{-1}\epsilon^{-6})$ and a communication complexity of $O(\epsilon^{-3})$ in terms of stationarity of f.
- To the best of our knowledge, we are the first to provide convergence results of general federated minimax problems with the PL-PL condition. We prove a better communication complexity of FESS-GDA in the PL-PL setting, compared with Local SGDA under the SC-SC setting (Deng and Mahdavi, 2021), despite that PL-PL is much weaker than SC-SC.

1.3 Related Works

Nonconvex-Strongly-Concave. For stochastic NC-SC problems, Lin et al. (2020) proved that SGDA achieves $O(\kappa^3\epsilon^{-4})$ sample complexity with a batch size of $O(\epsilon^{-2})$. Qiu et al. (2020); Luo et al. (2020) improved the sample complexity to $O(\kappa^3\epsilon^{-3})$ with a variance-reduction technique. Yang et al. (2022b) proved that Stochastic Smoothed-AGDA can achieve $O(\kappa^2\epsilon^{-4})$ sample complexity.

Nonconvex-Concave. Lin et al. (2020) analyzed GDA and SGDA for NC-C problems and proved that GDA can achieve $O(\epsilon^{-6})$ sample complexity for the deterministic setting and SGDA can achieve $O(\epsilon^{-8})$ sample complexity for the stochastic setting. Zhang et al. (2020) proposed Smoothed-AGDA and proved that it can achieve $O(\epsilon^{-4})$ sample complexity for the deterministic setting. For the stochastic setting, Rafique et al. (2021); Zhang et al. (2022) improved the complexity to $O(\epsilon^{-6})$ with a nested structure.

Federated minimax. There is a growing interest in solving federated minimax problems. Some focused

on the convex-concave setting (Deng et al., 2020a; Hou et al., 2021; Liao et al., 2021; Sun and Wei, 2022). There is also progress in the nonconvex setting. Deng et al. (2020b) analyzed a nonconvex-linear setting. Reisizadeh et al. (2020) formulated robust federated learning problems as special cases of federated PL-PL and NC-PL minimax problems and analyzed the convergence results of their proposed methods for these settings. Deng and Mahdavi (2021) proposed Local SGDA and Local SGDA+ and analyzed their convergence results under several nonconvex settings. Sharma et al. (2022) improved the convergence results in Deng and Mahdavi (2021). Yang et al. (2022a) proposed SAGDA and improved the communication complexity for the NC-PL setting. Sharma et al. (2023) proposed Fed-Norm-SGDA and Fed-Norm-SGDA+ and further improved the convergence results under several nonconvex settings. Tarzanagh et al. (2022) proposed FEDNEST with a nested structure and showed $O(\kappa^3\epsilon^{-4})$ sample complexity for the NC-SC setting. Huang (2022) designed AdaFGDA allowing for adaptive learning rates and improved the sample complexity to $\tilde{O}(\epsilon^{-3})$ for NC-PL setting with variance-reduction techniques. Recently, Wu et al. (2023) proposed FedSGDA-M and improved the sample complexity to $O(\kappa^3\epsilon^{-3})$ for the NC-PL setting with variance-reduction techniques.

2 PRELIMINARIES

Notations. We denote the l_2 norm as $\|\cdot\|_2$. For a differentiable function $g(x,y)$, we denote its gradient as $\nabla g(x,y) = (\nabla_x g(x,y)^T, \nabla_y g(x,y)^T)^T$. We define $\Phi(x) = \max_{y\in Y} f(x,y)$, $P_Y(y) = \text{argmin}_{y'\in Y} \frac{1}{2}\|y - y'\|^2$.

We state some common assumptions that will be used throughout the paper. They are commonly used in (federated) minimax optimization; e.g., (Yang et al., 2022b; Zhang et al., 2020; Deng and Mahdavi, 2021; Sharma et al., 2022).

Assumption 2.1 (Lipschitz smooth) *Each local function f_i is differentiable and there exists a positive constant l such that for all $i \in [M]$, and for all $x_1, x_2 \in X, y_1, y_2 \in Y$, we have*

$$\|\nabla f_i(x_1,y_1) - \nabla f_i(x_2,y_2)\| \le l[\|x_1 - x_2\| + \|y_1 - y_2\|].$$

Assumption 2.2 (Bounded variance) *The gradient of each local function $f_i(x,y,\xi_i)$, with a random data sample $\xi_i \sim \mathcal{D}_i$, is unbiased and has bounded variance, i.e., there exists a constant $\sigma > 0$ such that for all $i \in [M]$, and for all $x \in X, y \in Y$, $\mathbb{E}[\nabla f_i(x,y;\xi_i)] = \nabla f_i(x,y)$, and $\mathbb{E}\|\nabla f_i(x,y;\xi_i) - \nabla f_i(x,y)\|^2 \le \sigma^2$.*

Assumption 2.3 (Bounded heterogeneity) *To bound the heterogeneity of the local functions $\{f_i(x,y)\}$ across the clients, we assume there exits a constant $\sigma_G > 0$ such that*

$$\sup_{x\in X, y\in Y, i\in[M]} \|\nabla f_i(x,y) - \nabla f(x,y)\|^2 \le \sigma_G^2.$$

Assumption 2.4 *$\Phi(x) = \max_{y\in Y} f(x,y)$ is lower bounded by a finite $\Phi^* > -\infty$.*

The following notions of stationarity measures are also commonly used in the study of minimax optimization.

Definition 2.1 (Stationarity measures of f) *We say $(\hat{x},\hat{y})$ is an (ϵ_1,ϵ_2)-stationary point of a differentiable function $f(\cdot,\cdot)$ if $\|\nabla_x f(\hat{x},\hat{y})\| \le \epsilon_1$ and $l\|P_Y(\hat{y} + 1/l\nabla_y f(\hat{x},\hat{y})) - \hat{y}\| \le \epsilon_2$. If $(\hat{x},\hat{y})$ is an (ϵ,ϵ)-stationary point of f, we say it is an ϵ-stationary point of f.*

Definition 2.2 (Stationarity measures of Φ) *We say $\hat{x}$ is an ϵ-stationary point of a differentiable function $\Phi(\cdot)$ if $\|\nabla\Phi(\hat{x})\| \le \epsilon$.*

When f satisfies the PL condition in y, $\Phi(x) = \max_{y\in Y} f(x,y)$ is $2\kappa l$-Lipschitz smooth (Nouiehed et al., 2019). Thus, the stationarity measure of Φ is widely used in NC-PL and NC-SC settings. However, for other settings like NC-C, NC-1PC, $\Phi(x)$ is not guaranteed to be smooth, and the stationarity measure of the Moreau Envelope of the $\Phi(x)$ (Definition 2.4) is commonly used.

Definition 2.3 (Moreau envelope) *A function $\Phi_\lambda(z)$ is the Moreau envelope of $\Phi(x)$ with $\lambda > 0$, if for all $z \in X$, $\Phi_\lambda(z) = \min_x \Phi(x) + (1/2\lambda)\|z - x\|^2$.*

Definition 2.4 (Stationarity measures of $\Phi_{1/2l}$) *We say $\hat{x}$ is an ϵ-stationary point of $\Phi_{1/2l}(\cdot)$ if $\|\nabla\Phi_{1/2l}(\hat{x})\| \le \epsilon$.*

3 FESS-GDA

3.1 Algorithm

Inspired by the success of Smoothed-AGDA in the centralized setting (Zhang et al., 2020; Yang et al., 2022b), we propose `FESS-GDA`, which is compactly presented in Algorithm 1, for the federated minimax optimization problem. We consider a system with M clients and one central server. In each communication round, the server first randomly samples m clients and then sends them the current global model (x_t, y_t). For all participating clients, they synchronize their local models with the global model and perform K local updates with their local data and local learning rate $\eta_{x,l}, \eta_{y,l}$.

After the completion of local updates, each client sends back their local models to the server. Then, instead of a standard aggregation for local models like Local SGDA (Deng and Mahdavi, 2021), the key difference of FESS-GDA here is that we introduce an auxiliary parameter z_t to smooth the update of x_t.

Note that with a small local learning rate that $x_{t,i}^k \approx x_t, y_{t,i}^k \approx y_t$ and Assumption 2.2, the local updates can be approximated as

$$x_{t,i}^{k+1} \approx x_{t,i}^k - \eta_{x,l}\nabla_x f_i(x_t, y_t),$$
$$y_{t,i}^{k+1} \approx y_{t,i}^k + \eta_{y,l}\nabla_y f_i(x_t, y_t),$$

and with Assumption 2.3, the update of x_t, y_t can be approximated as

$$x_{t+1} \approx x_t - \eta_{x,l}\eta_{x,g}K[\nabla_x f(x_t, y_t) + p(x_t - z_t)],$$
$$y_{t+1} \approx y_t + \eta_{y,l}\eta_{y,g}K\nabla_y f(x_t, y_t),$$
$$z_{t+1} = z_t + \beta(x_{t+1} - z_t),$$

which has a similar form as the Smoothed-AGDA in the centralized setting.

Define $\hat{f}(x, y, z) = f(x, y) + \frac{p}{2}\|x - z\|^2$. Thus, in each communication round, we approximately perform gradient descent ascent of the following problem

$$\min_x \max_y \hat{f}(x, y, z_t) = f(x, y) + \frac{p}{2}\|x - z_t\|^2.$$

We set $\beta \in (0, 1)$ to guarantee that z_t is not too far from x_t. We choose $p = 2l$ for the NC-PL, NC-1PC, NC-C settings so that $\hat{f}(x, y, z)$ is l-strongly convex in x. For the PL-PL setting, since f satisfies the PL condition in x, we set $p = 0$. Note that when $p = 0, Y = \mathbb{R}^{d_2}$, FESS-GDA is equivalent to FSGDA (Yang et al., 2022a), and when $p = 0, Y = \mathbb{R}^{d_2}, \eta_{x,g} = \eta_{y,g} = 1$, FESS-GDA is equivalent to Local SGDA (Deng and Mahdavi, 2021).

Algorithm 1 FESS-GDA

Input: $x_0, y_0, z_0, \eta_{x,l}, \eta_{y,l}, \eta_{x,g}, \eta_{y,g}, \beta, p, T, K$
for $t = 0, 1, \cdots, T-1$ **do**
  Server randomly samples a subset S_t of clients with $|S_t| = m$, and send them (x_t, y_t).
  for each client $i \in S_t$ **do**
    $x_{t,i}^1 = x_t, y_{t,i}^1 = y_t$.
    for $k = 1, 2, \cdots, K$ **do**
      $x_{t,i}^{k+1} = x_{t,i}^k - \eta_{x,l}\nabla_x f_i(x_{t,i}^k, y_{t,i}^k, \xi_{t,i}^k)$
      $y_{t,i}^{k+1} = P_Y(y_{t,i}^k + \eta_{y,l}\nabla_y f_i(x_{t,i}^k, y_{t,i}^k, \xi_{t,i}^k))$
    end for
    Each client send their local models $(x_{t,i}^{K+1}, y_{t,i}^{K+1})$ to the server.
  end for
  Server aggregate local models and compute (x_{t+1}, y_{t+1}).
  $x_{t+1} = x_t + \eta_{x,g}(\frac{1}{m}\sum_{i\in S_t} x_{t,i}^{K+1} - x_t) - \eta_{x,l}\eta_{x,g}Kp(x_t - z_t)$
  $y_{t+1} = P_Y(y_t + \eta_{y,g}(\frac{1}{m}\sum_{i\in S_t} y_{t,i}^{K+1} - y_t))$
  $z_{t+1} = z_t + \beta(x_{t+1} - z_t)$
end for

3.2 Convergence

We analyze the convergence behaviors of FESS-GDA under the following settings. All proofs are deferred to the appendix.

3.2.1 Nonconvex-PL

Nonconvex-PL is a well-known weaker setting compared with Nonconvex-Strongly-Concave (NC-SC). Thus, the results in this section also hold for NC-SC.

Assumption 3.1 (PL condition in y) *Assume $X = \mathbb{R}^{d_1}, Y = \mathbb{R}^{d_2}$. For any fixed $x \in X$, $\max_{y\in Y} f(x, y)$ has a nonempty solution set and a finite optimal value. There exists $\mu > 0$ such that: $\|\nabla_y f(x, y)\|^2 \geq 2\mu[\max_y f(x, y) - f(x, y)], \forall x \in X, y \in Y$.*

We denote $\kappa := l/\mu$ in this section.

Theorem 3.1 *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 3.1, if we apply Algorithm 1 with appropriately chosen parameters (see Appendix D) and full client participation: $m = M$ or with homogeneous data: $\sigma_G = 0$, we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f with a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\kappa\epsilon^{-2})$. For partial client participation: $m < M$ and heterogeneous data: $\sigma_G > 0$, we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f with a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$.*

The formal statement and proof of Theorem 3.1 can be found in Appendix D. When $m = M$ or $\sigma_G = 0$, we set the number of local updates $K = \Theta(\kappa m^{-1}\epsilon^{-2})$ and can have a communication complexity of $O(\kappa\epsilon^{-2})$. However, when $m < M$ and $\sigma_G > 0$, our result does not show any convergence benefits from multiple local updates and can set $K = O(1)$ with a communication complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$. Similar behaviors have also been observed in other federated minimization and minimax works (Yang et al., 2021; Jhunjhunwala et al., 2022; Yang et al., 2022a; Sharma et al., 2023). As for the complexity, our per-client sample complexity exhibits a linear speedup w.r.t the number of participated clients.

When $M = 1$, our results recover the convergence results of Smoothed-AGDA in the centralized setting (Yang et al., 2022b). Similar to Yang et al. (2022b),

we can also translate an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f to an ϵ-stationary point of Φ under the federated setting, as stated below.

Proposition 3.1 (Translation) *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 3.1, if $(\tilde{x}, \tilde{y})$ is an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f, then we can find an $O(\epsilon)$-stationary point of Φ by solving $\min_x \max_y \{f(x,y) + l\|x - \tilde{x}\|^2\}$ from the initial point $(\tilde{x}, \tilde{y})$ using `FESS-GDA`. When $m = M$ or $\sigma_G = 0$, we need additional $O(\kappa^5 m^{-1}\epsilon^{-2}\log(\kappa))$ per-client sample complexity and $O(\kappa\log(\kappa))$ communication complexity. When $m < M$ and $\sigma_G > 0$, we need additional $O(\kappa^5 m^{-1}\epsilon^{-2}\log(\kappa))$ per-client sample complexity and $O(\kappa^5 m^{-1}\epsilon^{-2}\log(\kappa))$ communication complexity.*

With Proposition 3.1, we have the following corollary.

Corollary 3.1 *Under Assumptions 2.1, 2.2, 2.3, 2.4 and 3.1, when $m = M$ or $\sigma_G = 0$, we can use `FESS-GDA` to find an ϵ-stationary point of Φ with a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4} + \kappa^5 m^{-1}\epsilon^{-2}\log(\kappa))$ and a communication complexity of $O(\kappa\epsilon^{-2} + \kappa\log(\kappa))$. When $m < M$ and $\sigma_G > 0$, we can use `FESS-GDA` to find an ϵ-stationary point of Φ with a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4} + \kappa^5 m^{-1}\epsilon^{-2}\log(\kappa))$ and a communication complexity of $O(\kappa^2 m^{-1}\epsilon^{-4} + \kappa^5 m^{-1}\epsilon^{-2}\log(\kappa))$.*

When ϵ is small such that $\epsilon \leq \tilde{O}(\kappa^{-3/2})$, the sample and communication complexity needed to find an ϵ-stationary point of Φ have the same order as the complexity in finding $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f. Therefore, in terms of finding an ϵ-stationary point of Φ, our result presents the best-known communication complexity under similar settings. Compared with previous algorithms without variance reduction, we improve the sample complexity by a factor of $O(\kappa^2)$. We also establish additional convergence results in terms of stationarity of f.

3.2.2 Nonconvex-One-Point-Concave

Nonconvex-One-Point-Concave (Assumption 3.4) is a weaker setting than Nonconvex-Concave, and is studied in many federated minimax works (Deng and Mahdavi, 2021; Sharma et al., 2022, 2023). We use the following assumptions for this setting.

Assumption 3.2 (Compactness in y) *$X = \mathbb{R}^{d_1}$. Y is a convex, compact set of $\mathbb{R}^{d_2}$, and $D(Y)$ denotes the diameter of Y.*

Assumption 3.3 (Lipschitz continuity in y) *For any $x \in X, y, y' \in Y$, we have a finite number G_y, such that $\|f(x,y) - f(x,y')\| \leq G_y\|y - y'\|$.*

A similar assumption (Lipschitz continuity in x) is also used in Deng and Mahdavi (2021); Sharma et al. (2022, 2023).

Assumption 3.4 (One-Point-Concave in y) *For all $x \in X$, for all $y \in Y$, we have $\langle \nabla_y f(x,y), y - y^*(x)\rangle \leq f(x,y) - f(x,y^*(x))$, where $y^*(x) \in \operatorname{argmax}_{y\in Y} f(x,y)$.*

Theorem 3.2 *Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.2, 3.3, 3.4 and $\epsilon^2 \leq lD(Y)$, if we apply Algorithm 1 with appropriately chosen parameters (see Appendix F), with full client participation: $m = M$ or with homogeneous data: $\sigma_G = 0$, we can find an (ϵ, ϵ^2)-stationary point of f and an ϵ-stationary point of $\Phi_{1/2l}$ with a per-client sample complexity of $O(m^{-1}\epsilon^{-8})$ and a communication complexity of $O(\epsilon^{-4})$.*

We achieve comparable sample and communication complexity as the state-of-the-art algorithm Fed-Norm-SGDA+ (Sharma et al., 2023). However, their proof requires an additional assumption that each local loss function f_i satisfies the NC-1PC condition with a common global minimizer $y^*(x)$, while ours only requires the global loss functions to be NC-1PC. Moreover, several federated minimax works (Deng and Mahdavi, 2021; Sharma et al., 2022, 2023) assume $\|y_t\|^2 \leq D$, but did not specify how to guarantee it. We not only use this assumption but also use the projection operator in our algorithm to achieve this guarantee.

If we set $M = 1$, $K = 1$, and assume $\sigma = 0$, then Problem (1) reduces to the centralized deterministic minimax optimization problem and `FESS-GDA` reduces to Smoothed-GDA (Algorithm 2) (Zhang et al., 2020). Additionally, we have the following corollary for Smoothed-GDA under a centralized deterministic NC-1PC setting.

Algorithm 2 Smoothed-GDA

Input: $x_0, y_0, z_0, \eta_x, \eta_y, \beta, p, T$
for $t = 0, 1, ..., T-1$ **do**
  $x_{t+1} = x_t - \eta_x[\nabla_x f(x_t, y_t) + p(x_t - z_t)]$
  $y_{t+1} = P_Y(y_t + \eta_y \nabla_y f(x_t, y_t))$
  $z_{t+1} = z_t + \beta(x_{t+1} - z_t)$
end for

Corollary 3.2 *Under Assumptions 2.1, 2.4, 3.2, 3.3, 3.4, and when $M = 1$, $\epsilon^2 \leq lD(Y)$, if we apply Algorithm 2 with appropriately chosen parameters (see Appendix F), we can find an (ϵ, ϵ^2)-stationary point of f and an ϵ-stationary point of $\Phi_{1/2l}$ with a sample complexity of $O(\epsilon^{-4})$.*

Compared to the $O(\epsilon^{-4})$ sample complexity of Smoothed-GDA achieved in Zhang et al. (2020) under

NC-C setting, we achieve the same sample complexity under a weaker condition (NC-1PC).

3.2.3 Nonconvex-Concave

Since NC-1PC is weaker than NC-C, the results in Theorem 3.2 also hold for NC-C. Moreover, we have improved complexity results in terms of the stationarity of f, as presented in this section.

Assumption 3.5 (Concavity in y) *For all $x \in X$ and all $y, y' \in Y$, we have $f(x,y) \leq f(x,y') + \langle \nabla_y f(x,y'), y - y' \rangle$.*

Theorem 3.3 *Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.2, 3.3, 3.5 and $\epsilon \leq 2lD(Y)$, if we apply Algorithm 1 to optimize $\tilde{f}(x,y) = f(x,y) - \frac{\epsilon}{4D(Y)}\|y - y_0\|^2$, $y_0 \in Y$ with full client participation: $m = M$ or with homogeneous data: $\sigma_G = 0$, we can find an ϵ-stationary point of f with a per-client sample complexity of $O(m^{-1}\epsilon^{-6})$ and a communication complexity of $O(\epsilon^{-3})$.*

To the best of our knowledge, this is the best-known sample and communication complexity achieved in terms of stationarity of f under similar settings.

Moreover, we have the following corollary for the centralized deterministic setting.

Corollary 3.3 *Under Assumptions 2.1, 2.4, 3.2, 3.3, 3.5, when $M = 1$ and $\epsilon \leq 2lD(Y)$, we can apply Algorithm 2 to optimize $\tilde{f}(x,y) = f(x,y) - \frac{\epsilon}{4D(Y)}\|y - y_0\|^2$, $y_0 \in Y$, we can find an ϵ-stationary point of f with a sample complexity of $O(\epsilon^{-3})$.*

We improve the sample complexity of Smoothed-GDA under centralized deterministic NC-C setting from $O(\epsilon^{-4})$ to $O(\epsilon^{-3})$ in terms of stationarity of f.

3.2.4 Minimizing the Point-Wise Maximum of Finite Functions

We now consider optimizing f in a form of (2), which is widely used in practical applications. Zhang et al. (2020) proved that Smoothed-AGDA can achieve a sample complexity of $O(\epsilon^{-2})$ in terms of stationarity of f for solving (2) under centralized and deterministic settings, which is much better than the complexity needed for solving general nonconvex-concave problems. However, to the best of our knowledge, solving (2) under stochastic and federated settings remains unexplored.

For any stationary solution of (2) denoted as (x^*, y^*), the following KKT conditions hold:

$$\begin{aligned}
&\nabla F(x^*) y^* = 0, \\
&\sum_{i=1}^{m} y_i^* = 1, y_i^* \geq 0, \forall i \in [n], \\
&\lambda - \nu_i = f_i(x^*), \forall i \in [n], \nu_i \geq 0, \\
&\nu_i y_i^* = 0, \forall i \in [n],
\end{aligned}$$

where $\nabla F(x)$ denotes the Jacobian matrix of F at x, λ and ν are the multipliers for the equality constraint $\sum_{i=1}^{n} y_i = 1$, and the inequality constraint $y_i \geq 0$ respectively. We denote $\mathcal{I}_+(y^*)$ as the set of indices for which $y_i^* > 0$. We make following assumption on this set.

Assumption 3.6 (Strict complementarity) *For any stationary solution (x^*, y^*) of (2), we have $\nu_i > 0, \forall i \notin \mathcal{I}_+(y^*)$.*

Remark 3.1 *Assumption 3.6 is commonly used in the optimization literature (Forsgren et al., 2002; Carbonetto et al., 2009; Liang et al., 2014; Namkoong and Duchi, 2016; Lu et al., 2019; Zhang et al., 2020). This assumption generally holds if there is a linear term in the objective function and the data is from a continuous distribution (Zhang and Luo, 2020; Lu et al., 2019; Zhang et al., 2020).*

Theorem 3.4 *Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.3, 3.6, if we apply Algorithm 1 with appropriately chosen parameters (see Appendix H) to solve Problem (2), and assume $\|x_t\| \leq D_x$ for all t, then with full client participation: $m = M$ or with homogeneous data: $\sigma_G = 0$, we can find an ϵ-stationary point of f and $\Phi_{1/2l}$ with a per-client sample complexity of $O(m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\epsilon^{-2})$.*

To the best of our knowledge, we are the first to prove convergence results for solving Problem (2) under a federated setting. Setting $M = 1$, our results also indicate that we can find an ϵ-stationary point of f and $\Phi_{1/2l}$ of (2) with a sample complexity of $O(\epsilon^{-4})$ under the centralized stochastic setting. Assumptions similar to $\|x_t\| \leq D_x$ are also made in Deng and Mahdavi (2021); Sharma et al. (2022, 2023).

3.2.5 PL-PL

The PL-PL condition is much weaker than SC-SC and contains a richer class of functions. For example, according to Yang et al. (2020), $h(x,y) = x^2 + 3\sin^2 x \sin^2 y - 4y^2 - 10\sin^2 y$ satisfies Assumption 3.7, 3.8 (see Proposition 1 in Appendix of Yang et al. (2020)). However, $h(x,y)$ is nonconvex-nonconcave.

Reisizadeh et al. (2020) formulated robust federated learning as a special case of general federated minimax PL-PL problems and proposed FLRA. In their robust federated learning settings, each local client has its own local max variables and FLRA only communicates the min variables between the clients and the server. In this section, we consider a more general federated minimax setting (1) with the PL-PL condition. To the best of our knowledge, we are the first to prove convergence results for this general setting.

Assumption 3.7 (Two-sided PL condition) *Assume $X = \mathbb{R}^{d_1}, Y = \mathbb{R}^{d_2}$. For any fixed y, $\min_x f(x,y)$ has a nonempty solution set and a finite optimal value, and for any fixed x, $\max_y f(x,y)$ has a nonempty solution set and a finite optimal value. There exist constants $\mu_1, \mu_2 > 0$ such that: $\forall x, y$, $\|\nabla_x f(x,y)\|^2 \geq 2\mu_1[f(x,y) - \min_x f(x,y)]$ and $\|\nabla_y f(x,y)\|^2 \geq 2\mu_2[\max_y f(x,y) - f(x,y)]$.*

Assumption 3.8 (Existence of saddle point) *(x^*, y^*) is a saddle point of f if for any (x,y) : $f(x^*,y) \leq f(x^*,y^*) \leq f(x,y^*)$. We assume f has at least one saddle point.*

Since f is already μ_1-PL in x, we set $p = 0$ in this section. We further denote $\kappa' = \max\{l/\mu_1, l/\mu_2\}$, $\kappa'' = \min\{l/\mu_1, l/\mu_2\}$ in this section.

Theorem 3.5 *Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.7, 3.8, if we apply Algorithm 1 with appropriately chosen parameters (see Appendix I) for full client participation: $m = M$ or with homogeneous data: $\sigma_G = 0$, we can find (x_T, y_T) satisfying $\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq \epsilon^2$ with a per-client sample complexity of $O(m^{-1}\kappa'^3\kappa''^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$ and a communication complexity of $O(\kappa'\kappa''^2\log(\epsilon^{-1}\kappa'))$. For partial client participation: $m < M$ and heterogeneous data: $\sigma_G > 0$, we can find x_T, y_T satisfying $\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq \epsilon^2$ with a per-client sample complexity of $O(m^{-1}\kappa'^3\kappa''^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$ and a communication complexity of $O(m^{-1}\kappa'^3\kappa''^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$.*

When $M = 1$, our results recover the convergence results in Yang et al. (2020). For full client participation with heterogeneous data, we achieve a better communication complexity compared to Deng and Mahdavi (2021) and Reisizadeh et al. (2020). Moreover, we provide additional convergence results for partial client participation.

4 EXPERIMENTS

We perform GAN training and fair classification tasks in the federated setting to demonstrate the practical effectiveness and efficiency of `FESS-GDA` and verify our theoretical claims. We conduct our experiments on a computer with two NVIDIA RTX 3090 GPUs.

4.1 GAN

We consider a setting similar to Yang et al. (2022b), Loizou et al. (2020), using a Wasserstein GAN (Arjovsky et al., 2017) to approximate a one-dimensional Gaussian distribution in the federated setting. We first randomly generate a synthetic dataset of $n = 10000$ datapoints z sampled from a normal distribution with zero mean and unit variance and their corresponding real data $x^{\text{real}} = \hat{\mu} + \hat{\sigma}z$, where $\hat{\mu} = 0, \hat{\sigma} = 0.1$. We then evenly divide them into 10 disjoint sets for 10 clients. The generator is defined as $G_{\mu,\sigma}(z) = \mu + \sigma z$ and the discriminator is defined as $D_{\phi_1,\phi_2}(x) = \phi_1 x + \phi_2 x^2$. The problem can be formulated as

$$\min_{\mu,\sigma}\max_{\phi_1,\phi_2}\Big\{f(\mu,\sigma,\phi_1,\phi_2) = \frac{1}{n}\sum_{j=1}^{n} D_\phi(x_j^{\text{real}}) - D_\phi(G_{\mu,\sigma}(z_j)) - \lambda\|\phi\|^2\Big\},$$

where $\lambda > 0$ is the regularization coefficient to make the problem strongly concave.

We set a batch size of 100 for every update, and each client communicates with the server after every 10 local updates. We use the term $(\mu - \hat{\mu})^2 + (\sigma - \hat{\sigma})^2$ to measure the algorithm performances.

With $\lambda = 0.001, 0.005$ and 0.01, we compare the performances among Fed-Norm-SGDA, SAGDA and `FESS-GDA` (see Figure 1). We use $\beta = 0.05, p = 1$ for `FESS-GDA`. For each algorithm, we test their local learning rate from $\{1e-1, 1e-2, 1e-3\}$ and global learning rate from $\{1, 2\}$ in order to select the best for each algorithm under different λ. Each experiment is repeated 5 times and we report the average performance. As we can see from Figure 1, `FESS-GDA` achieves a significant speedup over Fed-Norm-SGDA and SAGDA with carefully tuned learning rates under different λ. Especially, when λ is relatively small, the performance gap between Fed-Norm-SGDA, SAGDA and `FESS-GDA` is more pronounced. Note that a smaller λ means a larger condition number κ (if we assume that the problem has a similar Lipschitz smooth constant l for different λ). This clearly validates our theoretical results that `FESS-GDA` improves the dependence of κ for nonconvex-strongly-concave problems.

4.2 Fair Classification

We consider a similar setting as Wu et al. (2023); Sharma et al. (2022); Nouiehed et al. (2019). The fair

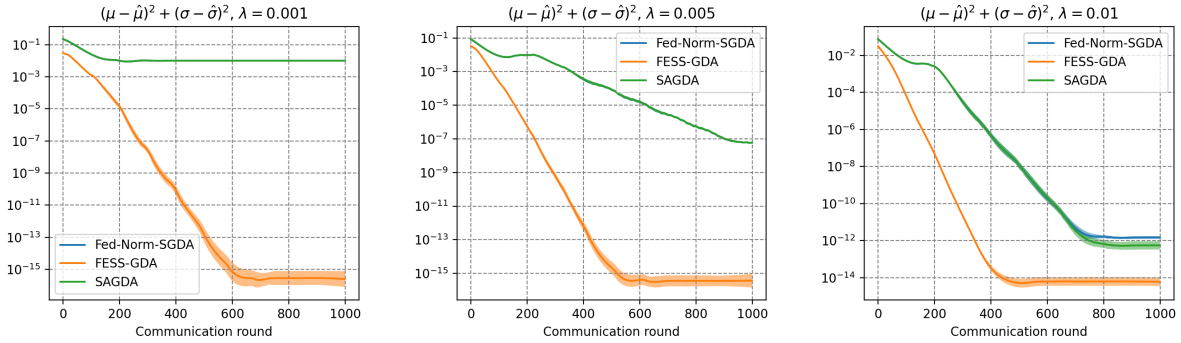

Figure 1: Comparison among Fed-Norm-SGDA, SAGDA and `FESS-GDA` for training a regularized WGAN with different regularization coefficients λ.

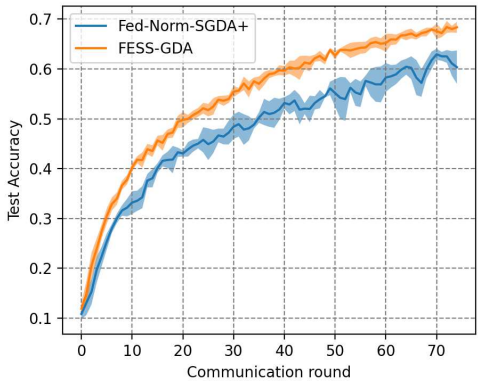

Figure 2: Comparison between Fed-Norm-SGDA+ and `FESS-GDA` for the fair classification task on CIFAR-10.

classification problem can be formulated as

$$\min_{x}\max_{y\in Y}\sum_{c=1}^{C}F_c(x)y_c,$$

where $Y=\{(y_1,...,y_C)^T|\sum_{c=1}^{C}y_c=1,y_c\geq 0\}$, x is the parameters of the model, and F_c is the loss function of class c. Clearly, this problem has the same form as (2) and is nonconvex-concave. We run the experiment on the CIFAR-10 dataset (Krizhevsky et al., 2009) with a convolutional neural network. We evenly divide the dataset into 10 disjoint sets for 10 clients. We compare the performances of Fed-Norm-SGDA+ and `FESS-GDA` for solving this problem and use the test accuracy as the performance measure. We set a batch size of 100 and inner loop $K=20$. For both algorithms, we adjust their local learning rate from $\{1e-1,1e-2\}$ and global learning rate from $\{1,1.5\}$. For `FESS-GDA`, we adjust its β from $\{0.1,0.5,0.9\}$ and its p from $\{0,1e-2,1e-1\}$. For Fed-Norm-SGDA+, we adjust its S from $\{1,5,10,20\}$. We tune all the parameters to achieve the best empirical performance for both algorithms. Each experiment is repeated 5 times and we report the average performance. As we can see from Figure 2, `FESS-GDA` achieves a better performance than Fed-Norm-SGDA+.

5 CONCLUSION

In this paper, we have proposed a new federated minimax optimization algorithm named `FESS-GDA`. We showed that `FESS-GDA` can be uniformly used for solving different classes of federated nonconvex minimax problems and theoretically established new or better convergence results for the considered settings. We further showcased the practical efficiency of `FESS-GDA` in practical federated learning tasks of training GANs and fair classification tasks.

Acknowledgements

The work of WS and CS was supported in part by the US National Science Foundation (NSF) under awards ECCS-2033671, ECCS-2143559, CPS-2313110, CNS-2002902, and Virginia Commonwealth Cyber Initiative Innovation and Commercialization Award VV-1Q23-005. The work of JZ was partially supported by MIT Postdoctoral Fellowship for Engineering Excellence 2023-2025.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Carbonetto, P., Schmidt, M., and Freitas, N. D. (2009). An interior-point stochastic approximation method and an l1-regularized delta rule. In *Advances in neural information processing systems*, pages 233–240.

Deng, Y., Kamani, M. M., and Mahdavi, M. (2020a). Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122.

Deng, Y., Kamani, M. M., and Mahdavi, M. (2020b). Distributionally robust federated averaging. In *Ad-*

vances in Neural Information Processing Systems, volume 33, pages 15111–15122.

Deng, Y. and Mahdavi, M. (2021). Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR.

Forsgren, A., Gill, P. E., and Wright, M. H. (2002). Interior methods for nonlinear optimization. *SIAM review*, 44(4):525–597.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hou, C., Thekumparampil, K. K., Fanti, G., and Oh, S. (2021). Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*.

Huang, F. (2022). Adaptive federated minimax optimization with lower complexities. *arXiv preprint arXiv:2211.07303*.

Jhunjhunwala, D., Sharma, P., Nagarkatti, A., and Joshi, G. (2022). Fedvarp: Tackling the variance due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pages 906–916. PMLR.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Léauté, T. and Faltings, B. (2013). Protecting privacy through distributed computation in multi-agent decision making. *Journal of Artificial Intelligence Research*, 47:649–695.

Liang, J., Fadili, J., and Peyré, G. (2014). Local linear convergence of forward–backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978.

Liao, L., Shen, L., Duan, J., Kolar, M., and Tao, D. (2021). Local adagrad-type algorithm for stochastic convex-concave minimax problems. *arXiv preprint arXiv:2106.10022*.

Lin, T., Jin, C., and Jordan, M. (2020). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR.

Liu, M., Yuan, Z., Ying, Y., and Yang, T. (2019). Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*.

Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. (2020). Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR.

Lu, S., Razaviyayn, M., Yang, B., Huang, K., and Hong, M. (2019). Snap: Finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems. *arXiv preprint arXiv:1907.04450*.

Luo, L., Ye, H., Huang, Z., and Zhang, T. (2020). Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 20566–20577.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems*, pages 2208–2216.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32.

Qiu, S., Yang, Z., Wei, X., Ye, J., and Wang, Z. (2020). Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear TD learning. *arXiv preprint arXiv:2008.10103*.

Rafique, H., Liu, M., Lin, Q., and Yang, T. (2021). Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. (2020). Robust federated learning: The case of affine distribution shifts. In *Advances in Neural Information Processing Systems*, volume 33, pages 21554–21565.

Sharma, P., Panda, R., and Joshi, G. (2023). Federated minimax optimization with client heterogeneity. *arXiv preprint arXiv:2302.04249*.

Sharma, P., Panda, R., Joshi, G., and Varshney, P. (2022). Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR.

Sun, Z. and Wei, E. (2022). A communication-efficient algorithm with linear convergence for federated minimax learning. *Advances in Neural Information Processing Systems*, 35:6060–6073.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. (2022). Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR.

Wu, X., Sun, J., Hu, Z., Zhang, A., and Huang, H. (2023). Solving a class of non-convex minimax optimization in federated learning. *arXiv preprint arXiv:2310.03613*.

Yang, H., Fang, M., and Liu, J. (2021). Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*.

Yang, H., Liu, Z., Zhang, X., and Liu, J. (2022a). Sagda: Achieving $\mathcal{O}(\epsilon^{-2})$ communication complexity in federated min-max learning. *Advances in Neural Information Processing Systems*, 35:7142–7154.

Yang, J., Kiyavash, N., and He, N. (2020). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165.

Yang, J., Orvieto, A., Lucchi, A., and He, N. (2022b). Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR.

Zhang, J. and Luo, Z.-Q. (2020). A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302.

Zhang, J., Xiao, P., Sun, R., and Luo, Z. (2020). A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389.

Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384.

Zhang, X., Aybat, N. S., and Gurbuzbalaban, M. (2022). Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *arXiv preprint arXiv:2205.15084*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Sections 2, 3.
   (b) Complete proofs of all theoretical results. [Yes] All proofs are in Appendix.
   (c) Clear explanations of any assumptions. [Yes] See Sections 2, 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 3.2.5.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Section 3.2.5.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Section 3.2.5.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] We used the CIFAR10 dataset and cited it.
   (b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material: Stochastic Smoothed Gradient Descent Ascent for Federated Minimax Optimization

The supplementary material is organized as follows. In Section A, we introduce notations that will be used throughout the supplementary material. In Section B, we present some preliminary lemmas. In Section C, we introduce necessary lemmas of the potential function for NC-PL and NC-1PC. In the subsequent sections, we provide the convergence results of `FESS-GDA` for NC-PL functions (Section D), NC-SC functions (Section E), NC-1PC functions (Section F), NC-C functions (Section G), functions having a form of (2) (Section H), and PL-PL functions (Section I). In Section J, we prove Proposition 3.1. Finally, in Section K, we provide additional results and details of our experiments.

A Notations

We introduce the following notations, which will play a significant role in our proof.

$$
\begin{aligned}
&\hat{f}(x,y,z) = f(x,y) + \frac{p}{2}\|x-z\|^2,\\
&\Psi(y,z) = \min_{x\in X} \hat{f}(x,y,z),\\
&\Phi(x) = \max_{y\in Y} f(x,y),\\
&\Phi(x,z) = \max_{y\in Y} \hat{f}(x,y,z),,\\
&P(z) = \min_{x\in X}\max_{y\in Y} \hat{f}(x,y,z),\\
&V_t = V(x_t,y_t,z_t) = \hat{f}(x_t,y_t,z_t) - 2\Psi(y_t,z_t) + 2P(z_t),\\
&x^*(y,z) = \arg\min_{x\in X} \hat{f}(x,y,z),\\
&x^*(z) = \arg\min_{x\in X} \Phi(x,z),\\
&y^*(x) \in \operatorname*{argmax}_{y\in Y} f(x,y),\\
&\hat{y}^*(z) \in \operatorname*{argmax}_{y\in Y} \Psi(y,z),\\
&x^+(y,z) = x - \eta_x K \nabla_x \hat{f}(x,y,z),\\
&y^+(z) = P_Y(y + \eta_y K \nabla_y f(x^*(y,z),y)).
\end{aligned}
$$

We denote $w_t = (x_t, y_t)$, $\eta_x = \eta_{x,g}\eta_{x,l}$, $\eta_y = \eta_{y,g}\eta_{y,l}$ for simplicity.

We summarize the main updates of `FESS-GDA` as:

$$
\begin{aligned}
&x_{t+1} = x_t - \eta_x K[u_{x,t} - e_{x,t} + p(x_t - z_t)],\\
&y_{t+1} = P_Y(y_t + \eta_y K(u_{y,t} - e_{y,t})),\\
&z_{t+1} = z_t + \beta(x_{t+1} - z_t),\\
&u_{x,t} = \frac{1}{m}\sum_{i\in S_t} \nabla_x f_i(w_t),\\
&u_{y,t} = \frac{1}{m}\sum_{i\in S_t} \nabla_y f_i(w_t),\\
&e_{x,t} = \frac{1}{mK}\sum_{i\in S_t}\sum_{j\in[K]} \left(\nabla_x f_i(w_t) - \nabla_x f_i(w_{t,i}^j, \xi_{t,i}^j)\right),
\end{aligned}
$$

$$\bar{e}_{x,t} = \mathbb{E}[e_{x,t}] = \frac{1}{mK}\sum_{i\in S_t}\sum_{j\in[K]}\left(\nabla_x f_i(w_t) - \nabla_x f_i(w_{t,i}^j)\right),$$
$$e_{y,t} = \frac{1}{mK}\sum_{i\in S_t}\sum_{j\in[K]}\left(\nabla_y f_i(w_t) - \nabla_y f_i(w_{t,i}^j, \xi_{t,i}^j)\right),$$
$$\bar{e}_{y,t} = \mathbb{E}[e_{y,t}] = \frac{1}{mK}\sum_{i\in S_t}\sum_{j\in[K]}\left(\nabla_y f_i(w_t) - \nabla_y f_i(w_{t,i}^j)\right).$$

We further define the following notations

$$d_{x,t} = \mathbb{E}\|\nabla_x \hat{f}(w_t, z_t) - u_{x,t} + e_{x,t} - p(x_t - z_t)\|^2,$$
$$d_{y,t} = \mathbb{E}\|\nabla_y f(w_t) - u_{y,t} + e_{y,t}\|^2.$$

Define $\bar{y}_{t+1} = P_Y(y_t + \eta_y K \nabla_y f(w_t))$, when $Y = \mathbb{R}^{d_2}$, we have $\bar{y}_{t+1} = y_t + \eta_y K \nabla_y f(w_t)$. Define $\Phi^* = \min_{x\in X}\max_{y\in Y} f(x,y)$, $\Delta = V_0 - \Phi^*$. Because $V(x,y,z) = P(z) + (f(x,y,z) - \Psi(y,z)) + (P(z) - \Psi(y,z)) \geq P(z) \geq \Phi^*$, we have

$$V_0 - V_t \leq V_0 - \min V_t \leq V_0 - \Phi^* = \Delta. \tag{4}$$

B Preliminary Lemmas

Lemma B.1 (Lemma C.1 (Yang et al., 2022b)) *When $p > l$, we have*

$$\|x^*(y,z) - x^*(y,z')\| \leq \gamma_1\|z - z'\|,$$
$$\|x^*(z) - x^*(z')\| \leq \gamma_1\|z - z'\|,$$
$$\|x^*(y,z) - x^*(y',z)\| \leq \gamma_2\|y - y'\|,$$

where $\gamma_1 = \frac{p}{-l+p}$, $\gamma_2 = \frac{l+p}{-l+p}$.

Lemma B.2 (Karimi et al. (2016)) *If function $g(x)$ is l-smooth and satisfies the PL condition with constant μ, then the following conditions hold*

$$g(x) - \min_z g(z) \geq \frac{\mu}{2}\|x_p - x\|^2,$$
$$\|\nabla_x g(x)\|^2 \geq 2\mu(g(x) - \min_z g(z)),$$
$$\|\nabla_x g(x)\|^2 \geq \mu\|x_p - x\|^2,$$

where x_p is the projection of x onto the optimal set.

Lemma B.3 *When $p = 2l$, we have*

$$\|\nabla_x \Phi(x^*(x_t))\| = \|\nabla_x \Phi_{1/2l}(x_t)\| = p\|x_t - x^*(x_t)\|.$$

Proof Note that $x^*(x_t) = \operatorname{argmin}_x\{\max_y f(x,y) + \frac{p}{2}\|x - x_t\|^2\}$. According to Lemma A.4 in Yang et al. (2022b), we have $\|\nabla_x \Phi(x^*(x_t))\| = \|\nabla_x \Phi_{1/2l}(x_t)\| = p\|x_t - x^*(x_t)\|$. ■

Lemma B.4 *When the local step sizes $\eta_{x,l}, \eta_{y,l}$ satisfy*

$$\eta_{x,l} \leq \frac{1}{2l\sqrt{2(2K-1)(K-1)}},$$
$$\eta_{y,l} \leq \frac{1}{2l\sqrt{2(2K-1)(K-1)}},$$

the following inequalities hold:

$$\mathbb{E}\|\bar{e}_{x,t}\|^2 \leq l^2[24K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f(w_t)\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 24K^2(\eta_{x,l}^2 + \eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2],$$
$$\mathbb{E}\|\bar{e}_{y,t}\|^2 \leq l^2[24K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f(w_t)\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 24K^2(\eta_{x,l}^2 + \eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2].$$

Proof According to the definition of $\|\bar{e}_{x,t}\|$, we have

$$
\begin{aligned}
&\mathbb{E}\|w_t - w_{t,i}^{j+1}\|^2 \\
=&\mathbb{E}\|x_{t,i}^j - \eta_{x,l}\nabla_x f_i(w_{t,i}^j, \xi_{t,i}^j) - x_t\|^2 + \mathbb{E}\|P_Y(y_{t,i}^j + \eta_{y,l}\nabla_y f_i(w_{t,i}^j, \xi_{t,i}^j)) - y_t\|^2 \\
\overset{(a)}{\leq}&\mathbb{E}\|x_{t,i}^j - x_t - \eta_{x,l}\nabla_x f_i(w_{t,i}^j)\|^2 + \mathbb{E}\|P_Y(y_{t,i}^j + \eta_{y,l}\nabla_y f_i(w_{t,i}^j, \xi_{t,i}^j)) - y_t\|^2 + \eta_{x,l}^2\sigma^2 \\
\overset{(b)}{\leq}&\left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|x_{t,i}^j - x_t\|^2 + 2K\eta_{x,l}^2\|\nabla_x f_i(w_{t,i}^j)\|^2 + \left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|y_{t,i}^j - y_t\|^2 + \\
&2K\mathbb{E}\|P_Y(y_{t,i}^j + \eta_{y,l}\nabla_y f_i(w_{t,i}^j, \xi_{t,i}^j)) - y_{t,i}^j\|^2 + \eta_{x,l}^2\sigma^2 \\
\overset{(c)}{\leq}&\left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|w_{t,i}^j - w_t\|^2 + 2K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_{t,i}^j)\|^2 + 2K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_{t,i}^j, \xi_{t,i}^j)\|^2 + \eta_{x,l}^2\sigma^2 \\
\overset{(d)}{\leq}&\left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|w_{t,i}^j - w_t\|^2 + 2K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_{t,i}^j)\|^2 + 2K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_{t,i}^j)\|^2 + (\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2 \\
\leq&\left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|w_{t,i}^j - w_t\|^2 + 4K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t)\|^2 + 4K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t)\|^2 + (\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2 + \\
&4K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t) - \nabla_x f_i(w_{t,i}^j)\|^2 + 4K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t) - \nabla_y f_i(w_{t,i}^j)\|^2 \\
\overset{(e)}{\leq}&\left(1 + \frac{1}{2K-1}\right)\mathbb{E}\|w_{t,i}^j - w_t\|^2 + 4K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t)\|^2 + 4K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t)\|^2 + (\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2 + \\
&4K\eta_{x,l}^2 l^2\|w_t - w_{t,i}^j\|^2 + 4K\eta_{y,l}^2 l^2\|w_t - w_{t,i}^j\|^2 \\
=&\left(1 + \frac{1}{2K-1} + 4Kl^2(\eta_{x,l}^2 + \eta_{y,l}^2)\right)\mathbb{E}\|w_{t,i}^j - w_t\|^2 + 4K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t)\|^2 + 4K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t)\|^2 + \\
&(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2 \\
\overset{(f)}{\leq}&\left(1 + \frac{1}{K-1}\right)\mathbb{E}\|w_{t,i}^j - w_t\|^2 + 4K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t)\|^2 + 4K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t)\|^2 + (\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2 \\
\overset{(g)}{\leq}&\sum_{\tau=0}^{j-1}\left(1 + \frac{1}{K-1}\right)^\tau [4K\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t)\|^2 + 4K\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t)\|^2 + (\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2] \\
\overset{(h)}{\leq}&12K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f_i(w_t)\|^2 + 12K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f_i(w_t)\|^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2 \\
\overset{(i)}{\leq}&24K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f(w_t)\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 24K^2(\eta_{x,l}^2 + \eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2.
\end{aligned}
$$

$(a), (d)$ are a consequence of the bounded variance of the stochastic gradient. (b) arises from the property that $\|a+b\|^2 \leq (1+c)\|a\|^2 + (1+1/c)\|b\|^2, c > 0$. (c) is a result of the nonexpansiveness of the projection operator. (e) is derived from the l-smoothness of the function f, while (f) is established based on the condition:

$$
\begin{aligned}
l^2\eta_{x,l}^2 \leq& \frac{1}{8(2K-1)(K-1)}, \\
l^2\eta_{y,l}^2 \leq& \frac{1}{8(2K-1)(K-1)}.
\end{aligned}
$$

(h) is due to

$$
\begin{aligned}
\sum_{\tau=0}^{j-1}\left(1 + \frac{1}{K-1}\right)^\tau \leq& (K-1)\left(\frac{K}{K-1}\right)^K \\
\frac{1}{K}\sum_{\tau=0}^{j-1}\left(1 + \frac{1}{K-1}\right)^\tau \leq& \left(\frac{K}{K-1}\right)^{K-1} \leq e \leq 3 \\
\sum_{\tau=0}^{j-1}\left(1 + \frac{1}{K-1}\right)^\tau \leq& 3K.
\end{aligned}
$$

(i) is from Assumption 2.3. We thus have

$$\mathbb{E}\|\bar{e}_{x,t}\|^2 \leq l^2[24K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f(w_t)\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].$$

The inequality for y can be proven in a similar fashion. ■

Lemma B.5 *The following inequalities establish upper bounds for $d_{x,t}$ and $d_{y,t}$:*

$$\begin{aligned}
d_{x,t} =&\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t) - u_{x,t} + e_{x,t} - p(x_t - z_t)\|^2 \leq 2\mathbb{E}\|\bar{e}_{x,t}\|^2 + 4\,(M-m)\,\frac{\sigma_G^2}{mM} + \frac{2}{mK}\sigma^2,\\
d_{y,t} =&\mathbb{E}\|\nabla_y f(w_t) - u_{y,t} + e_{y,t}\|^2 \leq 2\mathbb{E}\|\bar{e}_{y,t}\|^2 + 4\,(M-m)\,\frac{\sigma_G^2}{mM} + \frac{2}{mK}\sigma^2.
\end{aligned}$$

Proof According to the definition of $d_{x,t}$, we have

$$\begin{aligned}
d_{x,t} =&\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t) - u_{x,t} + e_{x,t} - p(x_t - z_t)\|^2\\
=&\mathbb{E}\|\nabla_x f(w_t) - u_{x,t} + e_{x,t}\|^2\\
\leq&2\mathbb{E}\|\nabla_x f(w_t) - u_{x,t}\|^2 + 2\mathbb{E}\|e_{x,t}\|^2\\
\leq&2\mathbb{E}\left\|\frac{1}{M}\sum_{j\in[M]}\nabla_x f_j(w_t) - \frac{1}{m}\sum_{i\in S_t}\nabla_x f_i(w_t)\right\|^2 + 2\mathbb{E}\|\bar{e}_{x,t}\|^2 + \frac{2}{mK}\sigma^2\\
\leq&2\mathbb{E}\left\|\left(\frac{1}{M}-\frac{1}{m}\right)\sum_{j\in S_t}\nabla_x f_j(w_t) - m\left(\frac{1}{M}-\frac{1}{m}\right)\nabla_x f(w_t) + \frac{1}{M}\sum_{i\in[M]/S_t}\nabla_x f_i(w_t) - \frac{M-m}{M}\nabla_x f(w_t)\right\|^2 +\\
&2\mathbb{E}\|\bar{e}_{x,t}\|^2 + \frac{2}{mK}\sigma^2\\
\leq&4\mathbb{E}\left\|\left(\frac{1}{M}-\frac{1}{m}\right)\sum_{j\in S_t}\nabla_x f_j(w_t) - m\left(\frac{1}{M}-\frac{1}{m}\right)\nabla_x f(w_t)\right\|^2 +\\
&4\mathbb{E}\left\|\frac{1}{M}\sum_{i\in[M]/S_t}\nabla_x f_i(w_t) - \frac{M-m}{M}\nabla_x f(w_t)\right\|^2 + 2\mathbb{E}\|\bar{e}_{x,t}\|^2 + \frac{2}{mK}\sigma^2\\
\overset{(a)}{\leq}&4\left[m\left(\frac{1}{M}-\frac{1}{m}\right)^2 + (M-m)\frac{1}{M^2}\right]\sigma_G^2 + 2\mathbb{E}\|\bar{e}_{x,t}\|^2 + \frac{2}{mK}\sigma^2\\
\leq&2\mathbb{E}\|\bar{e}_{x,t}\|^2 + 4\,(M-m)\,\frac{\sigma_G^2}{mM} + \frac{2}{mK}\sigma^2,
\end{aligned}$$

where (a) is due to $\|\sum_i^k x_i\|^2 \leq k\sum_i^k\|x_i\|^2$ and Assumption 2.3. Similarly, we have

$$d_{y,t} =\mathbb{E}\|\nabla_y f(w_t) - u_{y,t} + e_{y,t}\|^2 \leq 2\mathbb{E}\|\bar{e}_{y,t}\|^2 + 4\,(M-m)\,\frac{\sigma_G^2}{mM} + \frac{2}{mK}\sigma^2.$$

■

Lemma B.6 *Under the update rule of* `FESS-GDA`*, we have*

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_t\|^2 \leq&2\eta_x^2K^2\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 2\eta_x^2K^2 d_{x,t},\\
\mathbb{E}\|y_{t+1} - y_t\|^2 \leq&2\mathbb{E}\|\bar{y}_{t+1} - y_t\|^2 + 2\eta_y^2K^2 d_{y,t}.
\end{aligned}$$

When $Y = \mathbb{R}^{d_2}$, we have $\mathbb{E}\|y_{t+1} - y_t\|^2 \leq 2\eta_y^2K^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 2\eta_y^2K^2 d_{y,t}$.

Proof According to the update rule of x_t, y_t, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1}-x_t\|^2 =&\eta_x^2K^2\mathbb{E}\|u_{x,t}-e_{x,t}-p(x_t-z_t)\|^2\\
\leq&2\eta_x^2K^2\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+2\eta_x^2K^2\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)-u_{x,t}+e_{x,t}-p(x_t-z_t)\|^2\\
=&2\eta_x^2K^2\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+2\eta_x^2K^2d_{x,t},\\
\mathbb{E}\|y_{t+1}-y_t\|^2 =&\mathbb{E}\|P_Y(y_t+\eta_yK(u_{y,t}-e_{y,t}))-y_t\|^2\\
\overset{(a)}{\leq}&2\mathbb{E}\|P_Y(y_t+\eta_yK\nabla_yf(w_t))-y_t\|^2+2\eta_y^2K^2\|\nabla_yf(w_t)-u_{y,t}+e_{y,t}\|^2\\
=&2\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2+2\eta_y^2K^2d_{y,t},
\end{aligned}$$

where (a) is due to the nonexpansiveness of the projection operator. ■

C Intermediate Lemmas for Potential Function

Recall the potential function is defined as

$$V_t=V(x_t,y_t,z_t)=\hat{f}(x_t,y_t,z_t)-2\Psi(y_t,z_t)+2P(z_t).$$

The outline of the convergence proof for `FESS-GDA` aims to demonstrate the monotonic decrease of V_t. In this section, we present the essential lemmas required to establish bounds on the potential function.

Lemma C.1 *When $\eta_x\leq\frac{1}{4(p+l)K}$, we have the following inequality:*

$$\begin{aligned}
&\mathbb{E}\hat{f}(w_t,z_t)-\mathbb{E}\hat{f}(w_{t+1},z_t)\\
\geq&\frac{\eta_xK}{4}\|\nabla_x\hat{f}(w_t,z_t)\|^2-\frac{\eta_xK}{2}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}+\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2.
\end{aligned}$$

Proof Because of the $(p+l)$-smoothness of $\hat{f}(\cdot,z)$, we have

$$\begin{aligned}
&\mathbb{E}\hat{f}(w_t,z_t)-\mathbb{E}\hat{f}(w_{t+1},z_t)\\
\geq&\mathbb{E}\langle\nabla_x\hat{f}(w_t,z_t),x_t-x_{t+1}\rangle+\mathbb{E}\langle\nabla_y\hat{f}(w_t,z_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|x_{t+1}-x_t\|^2-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2\\
=&\eta_xK\mathbb{E}\langle\nabla_x\hat{f}(w_t,z_t),u_{x,t}+p(x_t-z_t)-\bar{e}_{x,t}\rangle+\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|x_{t+1}-x_t\|^2-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2\\
=&\eta_xK\mathbb{E}\langle\nabla_x\hat{f}(w_t,z_t),\nabla_x\hat{f}(w_t,z_t)-\bar{e}_{x,t}\rangle+\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|x_{t+1}-x_t\|^2-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2\\
=&\eta_xK\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{\eta_xK}{2}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)-\bar{e}_{x,t}\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2+\\
&\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|x_{t+1}-x_t\|^2-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2\\
\geq&\frac{\eta_xK}{2}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2+\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|x_{t+1}-x_t\|^2-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2\\
\overset{(a)}{\geq}&(\frac{\eta_xK}{2}-(p+l)\eta_x^2K^2)\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2-(p+l)\eta_x^2K^2d_{x,t}-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2+\\
&\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2\\
\overset{(b)}{\geq}&\frac{\eta_xK}{4}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}+\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2
\end{aligned}$$

where the (a) is due to the Lemma B.6, and (b) is due to the condition $\eta_x\leq\frac{1}{4(p+l)K}$. ■

Lemma C.2 *The z-update in `FESS-GDA` yields*

$$\hat{f}(w_{t+1},z_t)-\hat{f}(w_{t+1},z_{t+1})\geq\frac{p}{2\beta}\|z_t-z_{t+1}\|^2.$$

Proof By definition of $\hat{f}$ and the update rule of z, as $0 < \beta < 1$, we have

$$\begin{aligned}\hat{f}(w_{t+1}, z_t) - \hat{f}(w_{t+1}, z_{t+1}) =& \frac{p}{2}[\|x_{t+1} - z_t\|^2 - \|x_{t+1} - z_{t+1}\|^2] \\ =& \frac{p}{2}\left[\frac{1}{\beta^2}\|(z_{t+1} - z_t)\|^2 - \|(1-\beta)(x_{t+1} - z_t)\|^2\right] \\ =& \frac{p}{2}\left[\frac{1}{\beta^2}\|z_{t+1} - z_t\|^2 - \frac{(1-\beta)^2}{\beta^2}\|z_{t+1} - z_t\|^2\right] \\ \geq& \frac{p}{2\beta}\|z_t - z_{t+1}\|^2.\end{aligned}$$

■

Lemma C.3 *With $L_\Psi = l + l\gamma_2$, $\gamma_2 = \frac{l+p}{p-l}$, we have*

$$\begin{aligned}\Psi(y_{t+1}, z_t) - \Psi(y_t, z_t) \geq& \langle \nabla_y \Psi(y_t, z_t), y_{t+1} - y_t\rangle - \frac{L_\Psi}{2}\|y_{t+1} - y_t\|^2, \\ \Psi(y_{t+1}, z_{t+1}) - \Psi(y_{t+1}, z_t) \geq& \frac{p}{2}(z_{t+1} - z_t)^\top [z_{t+1} + z_t - 2x^*(y_{t+1}, z_{t+1})].\end{aligned}$$

Proof Since the dual function $\Psi(\cdot, z)$ is L_Ψ-smooth by Lemma B.3 in Zhang et al. (2020), we have

$$\Psi(y_{t+1}, z_t) - \Psi(y_t, z_t) \geq \langle \nabla_y \Psi(y_t, z_t), y_{t+1} - y_t\rangle - \frac{L_\Psi}{2}\|y_{t+1} - y_t\|^2.$$

By the definition of $x^*(y_{t+1}, z_t)$, we have

$$\begin{aligned}\Psi(y_{t+1}, z_{t+1}) - \Psi(y_{t+1}, z_t) =& \hat{f}(x^*(y_{t+1}, z_{t+1}), y_{t+1}, z_{t+1}) - \hat{f}(x^*(y_{t+1}, z_t), y_{t+1}, z_t) \\ \geq& \hat{f}(x^*(y_{t+1}, z_{t+1}), y_{t+1}, z_{t+1}) - \hat{f}(x^*(y_{t+1}, z_{t+1}), y_{t+1}, z_t) \\ =& \frac{p}{2}\left[\|z_{t+1} - x^*(y_{t+1}, z_{t+1})\|^2 - \|z_t - x^*(y_{t+1}, z_{t+1})\|^2\right] \\ =& \frac{p}{2}(z_{t+1} - z_t)^\top [z_{t+1} + z_t - 2x^*(y_{t+1}, z_{t+1})].\end{aligned}$$

■

Lemma C.4

$$P(z_{t+1}) - P(z_t) \leq \frac{p}{2}(z_{t+1} - z_t)^\top [z_{t+1} + z_t - 2x^*(\hat{y}^*(z_{t+1}), z_t)].$$

Proof By the definition of $\hat{y}^*(z_t)$ and $x^*(\hat{y}^*(z_{t+1}), z_{t+1})$, we have

$$\begin{aligned}P(z_{t+1}) - P(z_t) =& \Psi(\hat{y}^*(z_{t+1}), z_{t+1}) - \Psi(\hat{y}^*(z_t), z_t) \\ \leq& \Psi(\hat{y}^*(z_{t+1}), z_{t+1}) - \Psi(\hat{y}^*(z_{t+1}), z_t) \\ =& \hat{f}(x^*(\hat{y}^*(z_{t+1}), z_{t+1}), \hat{y}^*(z_{t+1}), z_{t+1}) - \hat{f}(x^*(\hat{y}^*(z_{t+1}), z_t), \hat{y}^*(z_{t+1}), z_t) \\ \leq& \hat{f}(x^*(\hat{y}^*(z_{t+1}), z_t), \hat{y}^*(z_{t+1}), z_{t+1}) - \hat{f}(x^*(\hat{y}^*(z_{t+1}), z_t), \hat{y}^*(z_{t+1}), z_t) \\ =& \frac{p}{2}(z_{t+1} - z_t)^\top [z_{t+1} + z_t - 2x^*(\hat{y}^*(z_{t+1}), z_t)].\end{aligned}$$

■

Lemma C.5 *The following inequality holds:*

$$\begin{aligned}&2\mathbb{E}(\Psi(y_{t+1}, z_{t+1}) - \Psi(y_{t+1}, z_t)) - 2\mathbb{E}(P(z_{t+1}) - P(z_t)) \\ \geq& -\left(2p\gamma_1 + \frac{p}{6\beta} - 48p\beta\gamma_1^2\right)\|z_{t+1} - z_t\|^2 - 24p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t, z_t)\|^2 - 48p\beta\gamma_2^2\|\bar{y}_{t+1} - y_t\|^2 - \\ &48p\beta\gamma_2^2\eta_y^2 K^2 d_{y,t}.\end{aligned}$$

Proof Combining Lemmas C.3 and C.4, we have

$$
\begin{aligned}
&2\mathbb{E}(\Psi(y_{t+1},z_{t+1})-\Psi(y_{t+1},z_t))-2\mathbb{E}(P(z_{t+1})-P(z_t))\\
\geq&2p\mathbb{E}(z_{t+1}-z_t)^\top[x^*(\hat{y}^*(z_{t+1}),z_t)-x^*(y_{t+1},z_{t+1})]\\
=&2p\mathbb{E}(z_{t+1}-z_t)^\top[x^*(\hat{y}^*(z_{t+1}),z_t)-x^*(\hat{y}^*(z_{t+1}),z_{t+1})]+2p\mathbb{E}(z_{t+1}-z_t)^\top[x^*(\hat{y}^*(z_{t+1}),z_{t+1})-x^*(y_{t+1},z_{t+1})]\\
\overset{(a)}{\geq}&-2p\gamma_1\|z_{t+1}-z_t\|^2+2p\mathbb{E}(z_{t+1}-z_t)^\top[x^*(\hat{y}^*(z_{t+1}),z_{t+1})-x^*(y_{t+1},z_{t+1})]\\
\geq&-\left(2p\gamma_1+\frac{p}{6\beta}\right)\|z_{t+1}-z_t\|^2-6p\beta\mathbb{E}\|x^*(z_{t+1})-x^*(y_{t+1},z_{t+1})\|^2\\
\geq&-\left(2p\gamma_1+\frac{p}{6\beta}\right)\|z_{t+1}-z_t\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(z_{t+1})\|^2-\\
&24p\beta\mathbb{E}\|x^*(y_{t+1},z_{t+1})-x^*(y_{t+1},z_t)\|^2-24p\beta\mathbb{E}\|x^*(y_{t+1},z_t)-x^*(y_t,z_t)\|^2\\
\overset{(b)}{\geq}&-\left(2p\gamma_1+\frac{p}{6\beta}+48p\beta\gamma_1^2\right)\|z_{t+1}-z_t\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2-24p\beta\gamma_2^2\|y_{t+1}-y_t\|^2\\
\overset{(c)}{\geq}&-\left(2p\gamma_1+\frac{p}{6\beta}+48p\beta\gamma_1^2\right)\|z_{t+1}-z_t\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2-48p\beta\gamma_2^2\|\bar{y}_{t+1}-y_t\|^2-\\
&48p\beta\gamma_2^2\eta_y^2K^2d_{y,t},
\end{aligned}
$$

where (a) and (b) are due to Lemma B.1, and (c) is due to Lemma B.6. ■

Lemma C.6 *Suppose we have $\eta_y \leq \frac{1}{8(2L_\Psi+l+p)K}$, $\eta_y=\eta_x/256$ and $p=2l$. In the unconstrained case when $Y=\mathbb{R}^{d_2}$, we have*

$$
\begin{aligned}
&\mathbb{E}\hat{f}(w_t,z_t)-\mathbb{E}\hat{f}(w_{t+1},z_t)+2(\mathbb{E}\Psi(y_{t+1},z_t)-\mathbb{E}\Psi(y_t,z_t))\\
\geq&\frac{\eta_xK}{8}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{\eta_yK}{8}\mathbb{E}\|\nabla_yf(w_t)\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}-\\
&\frac{3\eta_yK}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2-(2L_\Psi+l+p)\eta_y^2K^2d_{y,t}.
\end{aligned}
$$

In the constrained case when $Y\subset\mathbb{R}^{d_2}$ is convex and compact, we have

$$
\begin{aligned}
&\mathbb{E}\hat{f}(w_t,z_t)-\mathbb{E}\hat{f}(w_{t+1},z_t)+2(\mathbb{E}\Psi(y_{t+1},z_t)-\mathbb{E}\Psi(y_t,z_t))\\
\geq&\frac{\eta_xK}{8}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{8\eta_yK}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}-\\
&4\eta_yK\mathbb{E}\|\bar{e}_{y,t}\|^2-2\eta_yKd_{y,t},
\end{aligned}
$$

Proof Combining Lemmas C.2 and C.3, we have

$$
\begin{aligned}
&\mathbb{E}\hat{f}(w_t,z_t)-\mathbb{E}\hat{f}(w_{t+1},z_t)+2(\mathbb{E}\Psi(y_{t+1},z_t)-\mathbb{E}\Psi(y_t,z_t))\\
\geq&\frac{\eta_xK}{4}\|\nabla_x\hat{f}(w_t,z_t)\|^2-\frac{\eta_xK}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}+\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2+\\
&2\mathbb{E}\langle\nabla_y\Psi(y_t,z_t),y_{t+1}-y_t\rangle-L_\Psi\mathbb{E}\|y_{t+1}-y_t\|^2.
\end{aligned}
\tag{5}
$$

Denote $A_1=\mathbb{E}\langle\nabla_yf(w_t),y_t-y_{t+1}\rangle-\frac{p+l}{2}\mathbb{E}\|y_{t+1}-y_t\|^2+2\mathbb{E}\langle\nabla_y\Psi(y_t,z_t),y_{t+1}-y_t\rangle-L_\Psi\mathbb{E}\|y_{t+1}-y_t\|^2$.
When $Y=\mathbb{R}^{d_2}$, we have

$$
\begin{aligned}
A_1=&\mathbb{E}\langle\nabla_yf(w_t),y_{t+1}-y_t\rangle+2\mathbb{E}\langle\nabla_y\Psi(y_t,z_t)-\nabla_yf(w_t),y_{t+1}-y_t\rangle-\frac{2L_\Psi+l+p}{2}\mathbb{E}\|y_t-y_{t+1}\|^2\\
\overset{(a)}{\geq}&\eta_yK\mathbb{E}\langle\nabla_yf(w_t),u_{y,t}-e_{y,t}\rangle-2\eta_yK\mathbb{E}\langle\nabla_y\Psi(y_t,z_t)-\nabla_yf(w_t),u_{y,t}-e_{y,t}\rangle-\\
&\eta_y^2K^2(2L_\Psi+l+p)\|\nabla_yf(w_t)\|^2-\eta_y^2K^2(2L_\Psi+l+p)d_{y,t}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \eta_y K \mathbb{E}\langle \nabla_y f(w_t), \nabla_y f(w_t) - \bar{e}_{y,t}\rangle - 2\eta_y K \mathbb{E}\|\nabla_y \Psi(y_t, z_t) - \nabla_y f(w_t)\|\|\nabla_y f(w_t) - \bar{e}_{y,t}\| - \\
&\quad \eta_y^2 K^2 (2L_\Psi + l + p)\mathbb{E}\|\nabla_y f(w_t)\|^2 - \eta_y^2 K^2 (2L_\Psi + l + p) d_{y,t} \\
&\geq \eta_y K \mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t) - \bar{e}_{y,t}\|^2 - \\
&\quad 8\eta_y K \mathbb{E}\|\nabla_y \Psi(y_t, z_t) - \nabla_y f(w_t)\|^2 - \eta_y^2 K^2 (2L_\Psi + l + p)\mathbb{E}\|\nabla_y f(w_t)\|^2 - \eta_y^2 K^2 (2L_\Psi + l + p) d_{y,t} \\
&\geq \frac{\eta_y K}{2}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \frac{\eta_y K}{4}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \\
&\quad 8\eta_y K l^2 \mathbb{E}\|x^*(y_t, z_t) - x_t\|^2 - \eta_y^2 K^2 (2L_\Psi + l + p)\mathbb{E}\|\nabla_y f(w_t)\|^2 - \eta_y^2 K^2 (2L_\Psi + l + p) d_{y,t} \\
&\overset{(b)}{\geq} \left(\frac{\eta_y K}{4} - \eta_y^2 K^2 (2L_\Psi + l + p)\right)\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - (2L_\Psi + l + p)\eta_y^2 K^2 d_{y,t} - \\
&\quad \frac{8\eta_y K l^2}{(p-l)^2}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 \\
&\overset{(c)}{\geq} \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t)\|^2 - 8\eta_y K \mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - (2L_\Psi + l + p)\eta_y^2 K^2 d_{y,t}
\end{aligned} \tag{6}
$$

where (a) is due to Lemma B.6 and when $Y = \mathbb{R}^{d_2}, y_{t+1} - y_t = \eta_y K(u_{y,t} - e_{y,t})$, (b) is because of the $(p-l)$-strongly convexity of $\hat{f}(\cdot, y, z)$, and (c) is due to the condition $\eta_y \leq \frac{1}{8(2L_\Psi + l + p)K}$ and $p = 2l$. Combining (5) and (6), we have

$$
\begin{aligned}
&\mathbb{E}\hat{f}(w_t, z_t) - \mathbb{E}\hat{f}(w_{t+1}, z_t) + 2(\mathbb{E}\Psi(y_{t+1}, z_t) - \mathbb{E}\Psi(y_t, z_t)) \\
&\geq \left(\frac{\eta_x K}{4} - 8\eta_y K\right)\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - (p+l)\eta_x^2 K^2 d_{x,t} + \\
&\quad \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - (2L_\Psi + l + p)\eta_y^2 K^2 d_{y,t} \\
&\geq \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - (p+l)\eta_x^2 K^2 d_{x,t} - \\
&\quad \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - (2L_\Psi + l + p)\eta_y^2 K^2 d_{y,t},
\end{aligned}
$$

where the last inequality is because of the condition $\eta_y = \eta_x/256$.

When $Y \subset \mathbb{R}^{d_2}$ is convex and compact, we have

$$
\begin{aligned}
A_1 =& \mathbb{E}\langle \nabla_y f(w_t), y_{t+1} - y_t\rangle + 2\mathbb{E}\langle \nabla_y \Psi(y_t, z_t) - \nabla_y f(w_t), y_{t+1} - y_t\rangle - \frac{2L_\Psi + l + p}{2}\mathbb{E}\|y_t - y_{t+1}\|^2 \\
\geq& \mathbb{E}\langle u_{y,t} - e_{y,t}, y_{t+1} - y_t\rangle + \mathbb{E}\langle \nabla_y f(w_t) - u_{y,t} + e_{y,t}, y_{t+1} - \bar{y}_{t+1}\rangle + \mathbb{E}\langle \nabla_y f(w_t) - u_{y,t} + e_{y,t}, \bar{y}_{t+1} - y_t\rangle - \\
&2\mathbb{E}\|\nabla_y \Psi(y_t, z_t) - \nabla_y f(w_t)\|\|y_{t+1} - y_t\| - \frac{2L_\Psi + l + p}{2}\mathbb{E}\|y_t - y_{t+1}\|^2 \\
=& \mathbb{E}\langle u_{y,t} - e_{y,t}, P_Y(y_t + \eta_y K(u_{y,t} - e_{y,t})) - y_t\rangle + \\
&\mathbb{E}\langle \nabla_y f(w_t) - u_{y,t} + e_{y,t}, P_Y(y_t + \eta_y K(u_{y,t} - e_{y,t})) - P_Y(y_t + \eta_y K \nabla_y f(w_t))\rangle + \\
&\mathbb{E}\langle \bar{e}_{y,t}, \bar{y}_{t+1} - y_t\rangle - 2\mathbb{E}\|\nabla_y \Psi(y_t, z_t) - \nabla_y f(w_t)\|\|y_{t+1} - y_t\| - \frac{2L_\Psi + l + p}{2}\mathbb{E}\|y_t - y_{t+1}\|^2 \\
\geq& \frac{1}{\eta_y K}\mathbb{E}\|y_{t+1} - y_t\|^2 - \eta_y K \mathbb{E}\|\nabla_y f(w_t) - u_{y,t} + e_{y,t}\|^2 - 4\eta_y K \mathbb{E}\|\bar{e}_{y,t}\|^2 - \frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1} - y_t\|^2 - \\
&\frac{1}{8\eta_y K}\mathbb{E}\|y_{t+1} - y_t\|^2 - 8\eta_y K \mathbb{E}\|\nabla_y \Psi(y_t, z_t) - \nabla_y f(w_t)\|^2 - \frac{2L_\Psi + l + p}{2}\mathbb{E}\|y_t - y_{t+1}\|^2 \\
\geq& \left(\frac{1}{\eta_y K} - \frac{1}{8\eta_y K} - \frac{2L_\Psi + l + p}{2}\right)\mathbb{E}\|y_{t+1} - y_t\|^2 - \eta_y K d_{y,t} - 4\eta_y K \mathbb{E}\|\bar{e}_{y,t}\|^2 - \\
&\frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1} - y_t\|^2 - 8\eta_y K l^2 \mathbb{E}\|x_t - x^*(y_t, z_t)\|^2 \\
\overset{(a)}{\geq}& \frac{1}{2\eta_y K}\mathbb{E}\|y_{t+1} - y_t\|^2 - \eta_y K d_{y,t} - 4\eta_y K \mathbb{E}\|\bar{e}_{y,t}\|^2 - \frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1} - y_t\|^2 - \frac{8\eta_y K l^2}{(p-l)^2}\mathbb{E}\|\nabla_x f(w_t, z_t)\|^2
\end{aligned}
$$

$$\begin{aligned}
&\overset{(b)}{\geq}\frac{1}{4\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2-\frac{1}{\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_{t+1}\|^2-\eta_y K d_{y,t}-4\eta_y K\mathbb{E}\|\bar{e}_{y,t}\|^2-8\eta_y K\mathbb{E}\|\nabla_x f(w_t,z_t)\|^2-\\
&\frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2\\
&\geq\frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2-2\eta_y K d_{y,t}-4\eta_y K\mathbb{E}\|\bar{e}_{y,t}\|^2-8\eta_y K\mathbb{E}\|\nabla_x f(w_t,z_t)\|^2, \qquad (7)
\end{aligned}$$

where (a) is due to the condition $\eta_y \leq \frac{1}{8(2L_\Psi+l+p)K}$ and the $(p-l)$-strongly convexity of $\hat{f}(\cdot,y,z)$, (b) is due to the condition $p=2l$ and $\|a\|^2\geq\frac{1}{2}\|a-b\|^2-2\|b\|^2$. Combining (5) and (7), we have

$$\begin{aligned}
&\mathbb{E}\hat{f}(w_t,z_t)-\mathbb{E}\hat{f}(w_{t+1},z_t)+2(\mathbb{E}\Psi(y_{t+1},z_t)-\mathbb{E}\Psi(y_t,z_t))\\
&\geq\left(\frac{\eta_x K}{4}-8\eta_y K\right)\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2-\frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}+\\
&\frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2-4\eta_y K\mathbb{E}\|\bar{e}_{y,t}\|^2-2\eta_y K d_{y,t}\\
&\geq\frac{\eta_x K}{8}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2-\frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-(p+l)\eta_x^2K^2d_{x,t}-\\
&4\eta_y K\mathbb{E}\|\bar{e}_{y,t}\|^2-2\eta_y K d_{y,t},
\end{aligned}$$

where the last inequality is because of the condition $\eta_y=\eta_x/256$.

■

Lemma C.7 *Define potential function $V_t = V(x_t,y_t,z_t) = \hat{f}(x_t,y_t,z_t)-2\Psi(y_t,z_t)+2P(z_t)$, with $p=2l,\eta_x\leq 1/(1000Kl),\eta_y=\eta_x/256,\beta\leq\eta_y Kl/80000$, $\eta_{x,l}\leq\min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}},\sqrt{\frac{\beta}{6144\eta_x pl^2K^3}}\},\eta_{y,l}\leq\min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}},\sqrt{\frac{\eta_y}{3072\eta_x l^2K^2}}\}$, when $Y=\mathbb{R}^{d_2}$, we have*

$$\begin{aligned}
&\mathbb{E}V_t-\mathbb{E}V_{t+1}\\
&\geq\frac{\eta_x K}{32}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{\eta_y K}{32}\mathbb{E}\|\nabla_y f(w_t)\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-4\eta_x Kl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2];
\end{aligned}$$

when $Y\subset\mathbb{R}^{d_2}$ is convex and compact and under Assumption 3.3, we have

$$\begin{aligned}
&\mathbb{E}V_t-\mathbb{E}V_{t+1}\\
&\geq\frac{\eta_x K}{32}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{16\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_y K(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_x Kl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].
\end{aligned}$$

Proof Combining Lemma C.2, Lemma C.6 and Lemma C.5, when $Y=\mathbb{R}^{d_2}$, we have

$$\begin{aligned}
&\mathbb{E}V_t-\mathbb{E}V_{t+1}\\
&\geq\frac{\eta_x K}{8}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\left(\frac{\eta_y K}{8}-48p\beta\gamma_2^2\eta_y^2K^2\right)\mathbb{E}\|\nabla_y f(w_t)\|^2-24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2\\
&\left(\frac{p}{2\beta}-2p\gamma_1-\frac{p}{6\beta}-48p\beta\gamma_1^2\right)\mathbb{E}\|z_{t+1}-z_t\|^2-\frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2-\frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2-\\
&(p+l)\eta_x^2K^2d_{x,t}-\left(2L_\Psi+p+l+48p\beta\gamma_2^2\right)\eta_y^2K^2d_{y,t}
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p}{4\beta}\mathbb{E}\|z_{t+1}-z_t\|^2 - \\
&\quad 24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \\
&\quad (p+l)\eta_x^2 K^2 d_{x,t} - \left(2L_\Psi + p + l + 48p\beta\gamma_2^2\right)\eta_y^2 K^2 d_{y,t} \\
&\overset{(a)}{\geq} \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p\beta}{8}\mathbb{E}\|x_t-z_t\|^2 - \frac{p\beta}{2}\mathbb{E}\|x_{t+1}-x_t\|^2 - \\
&\quad 24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \\
&\quad (p+l)\eta_x^2 K^2 d_{x,t} - \left(2L_\Psi + p + l + 48p\beta\gamma_2^2\right)\eta_y^2 K^2 d_{y,t} \\
&\overset{(b)}{\geq} \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p\beta}{8}\mathbb{E}\|x_t-z_t\|^2 - p\beta\eta_x^2 K^2\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 - p\beta\eta_x^2 K^2 d_{x,t} \\
&\quad 24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \\
&\quad (p+l)\eta_x^2 K^2 d_{x,t} - \left(2L_\Psi + p + l + 48p\beta\gamma_2^2\right)\eta_y^2 K^2 d_{y,t} \\
&\geq \frac{\eta_x K}{16}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p\beta}{8}\mathbb{E}\|x_t-z_t\|^2 - 24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2 - \\
&\quad \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - (p+l+p\beta)\eta_x^2 K^2 d_{x,t} - \left(2L_\Psi + p + l + 48p\beta\gamma_2^2\right)\eta_y^2 K^2 d_{y,t}
\end{aligned}
$$

where in (a), we use

$$
\frac{p}{\beta}\|z_{t+1}-z_t\|^2 = p\beta\|x_{t+1}-z_t\|^2 \geq \frac{p\beta}{2}\|x_t-z_t\|^2 - 2p\beta\|x_{t+1}-x_t\|^2, \tag{8}
$$

and in (b), we use Lemma B.6.

Denote $A_2 = -\frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - \frac{3\eta_y K}{4}\mathbb{E}\|\bar{e}_{y,t}\|^2 - (p+l+p\beta)\eta_x^2 K^2 d_{x,t} - \left(2L_\Psi + p + l + 48p\beta\gamma_2^2\right)\eta_y^2 K^2 d_{y,t}$, we have

$$
\begin{aligned}
A_2 &\overset{(a)}{\geq} -(\eta_x K + 10l\eta_x^2 K^2)\mathbb{E}\|\bar{e}_{x,t}\|^2 - (\eta_y K + 24l\eta_y^2 K^2)\mathbb{E}\|\bar{e}_{y,t}\|^2 - \\
&\quad 20l\eta_x^2 K^2 (M-m)\frac{\sigma_G^2}{mM} - 10l\eta_x^2 K\frac{\sigma^2}{m} - 48l\eta_y^2 K^2 (M-m)\frac{\sigma_G^2}{mM} - 24l\eta_y^2 K\frac{\sigma^2}{m} \\
&\geq -25l\eta_x^2 K^2 (M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2 K\frac{\sigma^2}{m} - \\
&\quad 4\eta_x K l^2[24K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f(w_t)\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2] \\
&\overset{(b)}{\geq} -25l\eta_x^2 K^2 (M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2 K\frac{\sigma^2}{m} - 4\eta_x K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2] - \\
&\quad 4\eta_x K l^2[48K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + 48K^2p^2\eta_{x,l}^2\mathbb{E}\|x_t-z_t\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2] \\
&\overset{(c)}{\geq} -25l\eta_x^2 K^2 (M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2 K\frac{\sigma^2}{m} - 4\eta_x K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2] - \\
&\quad \frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 - \frac{\eta_y K}{32}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2,
\end{aligned}
$$

where (a) is because $(p+l+p\beta) \leq 5l, (2L_\Psi + p + l + 48p\beta\gamma_2^2) \leq 12l$ and Lemma B.5, in (b), we use Lemma B.4, in (c), we use the condition

$$
\begin{aligned}
\eta_{y,l}^2 &\leq \frac{\eta_y}{3072\eta_x l^2 K^2} \\
\eta_{x,l}^2 &\leq \frac{\beta}{6144\eta_x p l^2 K^3} \leq \frac{1}{6144 l^2 K^2}.
\end{aligned}
$$

So we have

$$
\mathbb{E}V_t - \mathbb{E}V_{t+1}
$$

$$\geq \frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{\eta_y K}{32}\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - 24p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t,z_t)\|^2 -$$
$$25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2K\frac{\sigma^2}{m} - 4\eta_x Kl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]. \tag{9}$$

When $Y \subset \mathbb{R}^{d_2}$ is convex and compact and under Assumption 3.3, similarly, we have

$$\begin{aligned}
&\mathbb{E}V_t - \mathbb{E}V_{t+1}\\
\geq&\frac{\eta_x K}{8}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \left(\frac{\eta_y K}{8} - 48p\beta\gamma_2^2\eta_y^2K^2\right)\frac{1}{\eta_y^2K^2}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2 - 24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2\\
&\left(\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta} - 48p\beta\gamma_1^2\right)\mathbb{E}\|z_{t+1}-z_t\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - 4\eta_y K\mathbb{E}\|\bar{e}_{y,t}\|^2 -\\
&(p+l)\eta_x^2K^2d_{x,t} - 2\eta_y K d_{y,t}\\
\geq&\frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{1}{16\eta_y K}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2 - 24p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t,z_t)\|^2 -\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2K\frac{\sigma^2}{m} - 8\eta_y K(M-m)\frac{\sigma_G^2}{mM} - \frac{4\eta_y\sigma^2}{m}\\
&4\eta_x Kl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2) + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].
\end{aligned} \tag{10}$$

The majority of the terms in (10) closely resemble those in (9). There are, however, two notable distinctions. First, there is an additional error term of $-8\eta_y K(M-m)\frac{\sigma_G^2}{mM} - \frac{4\eta_y\sigma^2}{m}$ attributed to the presence of $-2\eta_y K d_{y,t}$. Second, there is an additional error of $-96\eta_x K^3l^2G_y^2$, which arises from our utilization of Assumption 3.3. ■

D Nonconvex-PL

Lemma D.1 *Under Assumption 3.1 and $p = 2l$, we have*

$$\|x^*(y_t,z_t) - x^*(z_t)\|^2 \leq \frac{2}{l\mu}\|\nabla_y f(w_t)\|^2 + \frac{2}{l\mu}\|\nabla_x \hat{f}(w_t,z_t)\|^2.$$

Proof Because $\hat{f}(\cdot,y,z)$ is $(p-l)$-strongly convex, we have

$$\begin{aligned}
&\|x^*(y_t,z_t) - x^*(z_t)\|^2\\
\overset{(a)}{\leq}&\frac{2}{p-l}[\hat{f}(x^*(y_t,z_t),\hat{y}^*(z_t),z_t) - \hat{f}(x^*(z_t),\hat{y}^*(z_t),z_t)]\\
\leq&\frac{2}{p-l}[\Phi(x^*(y_t,z_t),z_t) - \Phi(x^*(z_t),z_t)]\\
=&\frac{2}{p-l}[\Phi(x^*(y_t,z_t),z_t) - \hat{f}(x^*(y_t,z_t),y_t,z_t) + \hat{f}(x^*(y_t,z_t),y_t,z_t) - \Phi(x^*(z_t),z_t)]\\
\overset{(b)}{\leq}&\frac{2}{p-l}[\Phi(x^*(y_t,z_t),z_t) - \hat{f}(x^*(y_t,z_t),y_t,z_t)]\\
\overset{(c)}{\leq}&\frac{1}{(p-l)\mu}\|\nabla_y f(x^*(y_t,z_t),y_t)\|^2\\
\overset{(d)}{\leq}&\frac{2}{(p-l)\mu}\|\nabla_y f(w_t)\|^2 + \frac{2}{(p-l)\mu}\|\nabla_y f(x^*(y_t,z_t),y_t) - \nabla_y f(x_t,y_t)\|^2\\
\leq&\frac{2}{(p-l)\mu}\|\nabla_y f(w_t)\|^2 + \frac{2l^2}{(p-l)\mu}\|x^*(y_t,z_t) - x_t\|^2\\
\overset{(e)}{\leq}&\frac{2}{(p-l)\mu}\|\nabla_y f(w_t)\|^2 + \frac{2l^2}{(p-l)^3\mu}\|\nabla_x \hat{f}(w_t,z_t)\|^2\\
=&\frac{2}{l\mu}\|\nabla_y f(w_t)\|^2 + \frac{2}{l\mu}\|\nabla_x \hat{f}(w_t,z_t)\|^2.
\end{aligned}$$

(b) can be attributed to the fact that $\hat{f}(x^*(y_t,z_t),y_t,z_t) \leq \Phi(x^*(z_t),z_t)$. (c) arises from the μ-PL property of $\hat{f}(x,\cdot,z)$. In $(a),(d),(e)$, we make use of Lemma B.2. ■

Proof of Theorem 3.1

We formally state Theorem 3.1 below.

Theorem 3.1 Under Assumptions 2.1, 2.2, 2.3, 2.4 and 3.1, with $p = 2l, \eta_y = \eta_x/256, \beta = \eta_y K\mu/80000$, $\eta_{x,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\beta}{6144\eta_x p l^2 K^3}}, O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}, \eta_{y,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\eta_y}{3072\eta_x l^2 K^2}}, O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, when $m = M$ or $\sigma_G = 0$, if we apply Algorithm 1 with $K = \Theta(\kappa m^{-1}\epsilon^{-2}), \eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}$, we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f with a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\kappa\epsilon^{-2})$; when $m < M$ and $\sigma_G > 0$, if we apply Algorithm 1 with $\eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{TlK}}\}, K = O(1)$, we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f with a per-client sample complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\kappa^2 m^{-1}\epsilon^{-4})$. Here, $\Delta = V_0 - \Phi^*$, $\kappa = l/\mu$.

Proof Combining Lemma D.1 and Lemma C.7, we have

$$\begin{aligned}\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq& \left(\frac{\eta_x K}{32} - \frac{96\beta}{\mu}\right)\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \left(\frac{\eta_y K}{32} - \frac{96\beta}{\mu}\right)\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - \\ & 25l\eta_x^2 K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2 K\frac{\sigma^2}{m} - 4\eta_x K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].\end{aligned}$$

Setting $\beta = \eta_y K\mu/80000$ yields

$$\begin{aligned}&\mathbb{E}V_t - \mathbb{E}V_{t+1} \\ \geq& \frac{\eta_x K}{64}\mathbb{E}\|\nabla_x \hat{f}(w_t)\|^2 + \frac{\eta_y K}{64}\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - \\ & 25l\eta_x^2 K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2 K\frac{\sigma^2}{m} - 4\eta_x K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]. \qquad (11)\end{aligned}$$

Further note that

$$\|\nabla_x f(x_t,y_t)\|^2 \leq 2\|\nabla_x \hat{f}(w_t,z_t)\|^2 + 2p^2\|x_t - z_t\|^2, \qquad (12)$$

which leads to

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 + \kappa\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 \\ \leq& \frac{1}{T}\sum_{t=0}^{T-1}\max\left\{\frac{128\kappa}{\eta_x K}, \frac{64\kappa}{\eta_y K}, \frac{32p}{\beta}\right\}\Big\{\mathbb{E}V_t - \mathbb{E}V_{t+1} + \\ & 25l\eta_x^2 K^2(M-m)\frac{\sigma_G^2}{mM} + 15l\eta_x^2 K\frac{\sigma^2}{m} + 4\eta_x K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]\Big\} \\ \overset{(a)}{\leq}& \frac{O(1)\kappa}{\eta_x KT}[V_0 - \min_t V_t] + O(1)\kappa\eta_x lK(M-m)\frac{\sigma_G^2}{mM} + O(1)\kappa\eta_x l\frac{\sigma^2}{m} + \\ & O(1)\kappa l^2[K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2] \\ \leq& O(1)\frac{\kappa\Delta}{\eta_x KT} + O(1)\kappa\eta_x lK(M-m)\frac{\sigma_G^2}{mM} + O(1)\kappa\eta_x l\frac{\sigma^2}{m} + \\ & O(1)\kappa l^2[K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2], \qquad (13)\end{aligned}$$

where (a) is because $p/\beta = O(1)\kappa/(\eta_x K)$ and (4).

When $m = M$ or $\sigma_G = 0$, with $\eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}$, $\eta_{x,l} \leq O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1}), \eta_{y,l} \leq O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1})$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 + \kappa\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 \leq O(1)\frac{\kappa}{\sqrt{mKT}} + O(1)\frac{\kappa}{T} + O(1)\epsilon^2,$$

which implies that we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f with a per-client sample complexity of $KT = O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $T = O(\kappa\epsilon^{-2})$.

When $m < M$ and $\sigma_G > 0$, with $\eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{Tl}K}\}$, $K = O(1)$, $\eta_{x,l} \leq O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1}), \eta_{y,l} \leq O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1})$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 + \kappa\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 \leq O(1)\frac{\kappa}{\sqrt{mT}} + O(1)\epsilon^2,$$

which implies that we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f with a per-client sample complexity of $KT = O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $T = O(\kappa^2 m^{-1}\epsilon^{-4})$.

■

E Nonconvex-Strongly-Concave

Since Nonconvex-PL is weaker than Nonconvex-Strongly-Concave (NC-SC), Theorem 3.1 also holds for NC-SC. However, for NC-SC, Theorem E.1 proves that `FESS-GDA` can achieve similar convergence results when Y is a convex, compact set of $\mathbb{R}^{d_2}$.

Assumption E.1 (Strongly Concave in y) *f is μ-strongly concave ($\mu > 0$) in y, if for any fixed x, $\max_y f(x,y)$, $\forall y, y' \in Y$, we have*

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

Lemma E.1 *Define $y_+(x) = P_Y(y+\eta_y K\nabla_y f(x,y))$. Under Assumptions 3.1, 3.2, 2.1, and with $\eta_y K \leq 1/1000l$, we have*

$$\|y - y^*(x)\| \leq \frac{2}{\mu\eta_y K}\|y - y_+(x)\|,$$
$$\|y_+(x) - y^*(x)\| \leq \frac{2}{\mu\eta_y K}\|y - y_+(x)\|.$$

Proof We define $\hat{g}(y;v) = \|y\|^2 - 2\langle y, v\rangle + \mathbf{1}_Y(y), v_1 = y + \eta_y K\nabla_y f(x,y), v_2 = y^*(x) + \eta_y K\nabla_y f(x, y^*(x))$. According to the definition of $y_+(x), y^*(x)$, we have

$$y_+(x) = \arg\min_y \hat{g}(y;v_1)$$
$$y^*(x) = \arg\min_y \hat{g}(y;v_2).$$

Note that $\hat{g}(\cdot;v)$ is 2-strongly-convex, according to Lemma B.2, we have

$$\hat{g}(y_+(x);v_2) - \hat{g}(y^*(x);v_2) \geq \|y_+(x) - y^*(x)\|^2 \tag{14}$$
$$\hat{g}(y^*(x);v_1) - \hat{g}(y_+(x);v_1) \geq \|y_+(x) - y^*(x)\|^2. \tag{15}$$

By the definition of $\hat{g}$:

$$\hat{g}(y_+(x);v_1) - \hat{g}(y_+(x);v_2) = -2\langle y_+(x), v_1 - v_2\rangle, \tag{16}$$

$$\hat{g}(y^*(x);v_1)-\hat{g}(y^*(x);v_2)=-2\langle y^*(x),v_1-v_2\rangle. \tag{17}$$

Combining (14),(15),(16),(17), we have

$$\|y_+(x)-y^*(x)\|^2\le\langle y_+(x)-y^*(x),v_1-v_2\rangle. \tag{18}$$

Therefore,

$$\|y_+(x)-y^*(x)\|\le\|v_1-v_2\|. \tag{19}$$

By the definition of v_1,v_2, we have

$$\begin{aligned}\|v_1-v_2\|^2=&\|y-y^*(x)\|^2+2\eta_yK\langle y-y^*(x),\nabla_yf(x,y)-\nabla_yf(x,y^*(x))\rangle+\eta_y^2K^2\|\nabla_yf(x,y)-\nabla_yf(x,y^*(x))\|^2\\ \overset{(a)}{\le}&\|y-y^*(x)\|^2+(2\eta_yK-\eta_y^2K^2l)\langle y-y^*(x),\nabla_yf(x,y)-\nabla_yf(x,y^*(x))\rangle\\ \overset{(b)}{\le}&\|y-y^*(x)\|^2+\eta_yK\langle y-y^*(x),\nabla_yf(x,y)-\nabla_yf(x,y^*(x))\rangle\\ \overset{(c)}{\le}&(1-\eta_yK\mu)\|y-y^*(x)\|^2\\ \le&\left(1-\frac{\eta_yK\mu}{2}\right)^2\|y-y^*(x)\|^2\end{aligned} \tag{20}$$

where (a) is a consequence of several factors. Firstly, due to the concavity of f in y, we have $\langle y-y^*(x),\nabla_yf(x,y)-\nabla_yf(x,y^*(x))\rangle\le0$. Additionally, Assumption 2.1 ensures that $\|\nabla_yf(x,y)-\nabla_yf(x,y^*(x))\|\le l\|y-y^*(x)\|$. (b) follows from the condition $\eta_yK\le1/l$, and (c) stems from the μ-strong concavity of $f(x,\cdot)$.

Combining (19), (20), we have

$$\|y_+(x)-y^*(x)\|\le\|v_1-v_2\|\le(1-\eta_yK\mu/2)\|y-y^*(x)\|.$$

So

$$\|y-y_+(x)\|\ge\|y-y^*(x)\|-\|y_+(x)-y^*(x)\|\ge\frac{\eta_yK\mu}{2}\|y-y^*(x)\|,$$

$$\|y-y^*(x)\|\le\frac{2}{\eta_yK\mu}\|y-y_+(x)\|,$$

which yields

$$\|y_+(x)-y^*(x)\|\le(1-\eta_yK\mu/2)\|y-y^*(x)\|\le\frac{2}{\eta_yK\mu}\|y-y_+(x)\|.$$

■

Lemma E.2 *Under Assumption 3.1, 3.2, 2.1, and with $\eta_yK\le1/1000l$, the following inequelity holds*

$$\hat{f}(x^*(y,z),y^*(x^*(y,z)),z)-\hat{f}(x^*(y,z),y^+(z),z)\le\frac{2(1+\eta_yKl)}{\mu\eta_y^2K^2}\|y-y^+(z)\|^2.$$

Proof Noting that $\hat{f}$ is μ-strongly concave in y, we have

$$\begin{aligned}&\hat{f}(x^*(y,z),y^*(x^*(y,z)),z)-\hat{f}(x^*(y,z),y^+(z),z)\\ \le&\langle\nabla_y\hat{f}(x^*(y,z),y^+(z),z),y^*(x^*(y,z))-y^+(z)\rangle-\frac{\mu}{2}\|y^*(x^*(y,z))-y^+(z)\|\\ \le&\langle\nabla_y\hat{f}(x^*(y,z),y^+(z),z),y^*(x^*(y,z))-y^+(z)\rangle\\ =&\langle\nabla_y\hat{f}(x^*(y,z),y,z),y^*(x^*(y,z))-y^+(z)\rangle+\end{aligned}$$

$$
\begin{aligned}
&\langle \nabla_y \hat{f}(x^*(y,z),y^+(z),z) - \nabla_y \hat{f}(x^*(y,z),y,z), y^*(x^*(y,z)) - y^+(z)\rangle \\
\overset{(a)}{\leq} &\frac{1}{\eta_y K}\langle y^+(z) - y, y^*(x^*(y,z)) - y^+(z)\rangle + \\
&\frac{1}{\eta_y K}\langle y + \eta_y K \nabla_y \hat{f}(x^*(y,z),y,z) - y^+(z), y^*(x^*(y,z)) - y^+(z)\rangle + \\
&l\|y - y^+(z)\|\|y^*(x^*(y,z)) - y^+(z)\| \\
\overset{(b)}{\leq} &\frac{1+\eta_y K l}{\eta_y K}\|y - y^+(z)\|\|y^*(x^*(y,z)) - y^+(z)\| \\
\overset{(c)}{\leq} &\frac{2(1+\eta_y K l)}{\mu \eta_y^2 K^2}\|y - y^+(z)\|^2,
\end{aligned}
$$

where in (a), we use the l-smoothness of f and $\nabla_y \hat{f} = \nabla_y f$, in (b), we use the fact that when Y is a closed convex set, we have

$$
\langle a - P_Y(a), b - P_Y(a)\rangle \leq 0, \quad \forall b \in Y, \tag{21}
$$

and in (c), we use Lemma E.1. ■

Lemma E.3 *Under Assumption 3.1, 3.2, 2.1, and with $\eta_y K \leq 1/1000l$, we have*

$$
\begin{aligned}
&\|x^*(y_t,z_t) - x^*(z_t)\|^2 \\
\leq &\frac{10}{\mu \eta_y^2 K^2 l}\|y_t - \bar{y}_{t+1}\|^2 + \frac{10}{\mu l}\|\nabla_x \hat{f}(x_t,y_t,z_t)\|^2 + \frac{40}{l^2}\|\nabla_x \hat{f}(x_t,y_t,z_t)\|^2.
\end{aligned}
$$

Proof Noting that since $\hat{f}(\cdot,y,z)$ is $(l+p)$-smooth, we have

$$
\begin{aligned}
&\hat{f}(x^*(y,z),y^+(z),z) - \hat{f}(x^*(y^+(z),z),y^+(z),z) \\
\leq &\langle \nabla_x \hat{f}(x^*(y^+(z),z),y^+(z),z), x^*(y,z) - x^*(y^+(z),z)\rangle + \frac{l+p}{2}\|x^*(y,z) - x^*(y^+(z),z)\|^2 \\
\leq &\frac{1}{2l}\|\nabla_x \hat{f}(x^*(y^+(z),z),y^+(z),z)\|^2 + \frac{2l+p}{2}\|x^*(y,z) - x^*(y^+(z),z)\|^2 \\
\leq &\frac{2}{l}\|\nabla_x \hat{f}(x,y,z)\|^2 + \frac{2}{l}\|\nabla_x \hat{f}(x,y,z) - \nabla_x \hat{f}(x^*(y,z),y,z)\|^2 + \frac{2}{l}\|\nabla_x \hat{f}(x^*(y,z),y,z) - \nabla_x \hat{f}(x^*(y^+(z),z),y,z)\|^2 + \\
&\frac{2}{l}\|\nabla_x \hat{f}(x^*(y^+(z),z),y,z) - \nabla_x \hat{f}(x^*(y^+(z),z),y^+(z),z)\|^2 + \frac{2l+p}{2}\|x^*(y,z) - x^*(y^+(z),z)\|^2 \\
\leq &\frac{2}{l}\|\nabla_x \hat{f}(x,y,z)\|^2 + \frac{2(p+l)^2}{l}\|x - x^*(y,z)\|^2 + \left(\frac{2(p+l)^2}{l} + \frac{2l+p}{2}\right)\|x^*(y,z) - x^*(y^+(z),z)\|^2 + 2l\|y - y^+(z)\|^2 \\
\overset{(a)}{\leq} &\frac{20}{l}\|\nabla_x \hat{f}(x,y,z)\|^2 + (20\gamma_2 + 2)l\|y - y^+(z)\|^2,
\end{aligned} \tag{22}
$$

where we use strong convexity of $\hat{f}(\cdot,y,z)$ and Lemma B.1 to establish (a). By the strong convexity of $\hat{f}(\cdot,y,z)$, we have

$$
\begin{aligned}
&\|x^*(y,z) - x^*(z)\|^2 \\
\leq &\frac{2}{p-l}[\hat{f}(x^*(y,z),\hat{y}^*(z),z) - \hat{f}(x^*(z),\hat{y}^*(z),z)] \\
\overset{(a)}{\leq} &\frac{2}{p-l}[\Phi(x^*(y,z),z) - \hat{f}(x^*(y^+(z),z),y^+(z),z) + \hat{f}(x^*(y^+(z),z),y^+(z),z) - \Phi(x^*(z),z)] \\
\overset{(b)}{\leq} &\frac{2}{p-l}[\hat{f}(x^*(y,z),y^*(x^*(y,z)),z) - \hat{f}(x^*(y^+(z),z),y^+(z),z)] \\
\overset{(c)}{\leq} &\frac{2}{l}[\hat{f}(x^*(y,z),y^*(x^*(y,z)),z) - \hat{f}(x^*(y,z),y^+(z),z)] + \frac{40}{l^2}\|\nabla_x \hat{f}(x,y,z)\|^2 + (40\gamma_2 + 4)\|y - y^+(z)\|^2
\end{aligned}
$$

$$\begin{aligned}
&\overset{(d)}{\leq}\frac{4(1+\eta_y Kl)+(40\gamma_2+4)\mu l\eta_y^2K^2}{\mu\eta_y^2K^2l}\|y-y^+(z)\|^2+\frac{40}{l^2}\|\nabla_x\hat{f}(x,y,z)\|^2\\
&\leq\frac{5}{\mu\eta_y^2K^2l}\|y-y^+(z)\|^2+\frac{40}{l^2}\|\nabla_x\hat{f}(x,y,z)\|^2,
\end{aligned}$$

where (a) is because that $\hat{f}(x^*(y,z),\hat{y}^*(z),z)\leq\Phi(x^*(y,z),z)$, $\Phi(x^*(z),z)=\hat{f}(x^*(z),\hat{y}^*(z),z)$, (b) is because $\hat{f}(x^*(y^+(z),z),y^+(z),z)\leq\Phi(x^*(z),z)$, (c) is due to (22), and (d) is due to Lemma E.2. Then, we have

$$\begin{aligned}
&\|x^*(y_t,z_t)-x^*(z_t)\|^2\\
\leq&\frac{5}{\mu\eta_y^2K^2l}\|y_t-y_t^+(z_t)\|^2+\frac{40}{l^2}\|\nabla_x\hat{f}(x_t,y_t,z_t)\|^2\\
\overset{(a)}{\leq}&\frac{10}{\mu\eta_y^2K^2l}\|y_t-\bar{y}_{t+1}\|^2+\frac{10l}{\mu}\|x_t-x^*(y_t,z_t)\|^2+\frac{40}{l^2}\|\nabla_x\hat{f}(x_t,y_t,z_t)\|^2\\
\overset{(b)}{\leq}&\frac{10}{\mu\eta_y^2K^2l}\|y_t-\bar{y}_{t+1}\|^2+\frac{10}{\mu l}\|\nabla_x\hat{f}(x_t,y_t,z_t)\|^2+\frac{40}{l^2}\|\nabla_x\hat{f}(x_t,y_t,z_t)\|^2,
\end{aligned}$$

where in (a), we use l-smoothness of f and in (b), we use strong convexity of $\hat{f}(\cdot,y,z)$. ■

Theorem E.1 *Under the Assumptions 2.1, 2.2, 2.3, 2.4, 3.2, 3.3, E.1, if we apply Algorithm 1 with* $p=2l$, $\eta_x=\min\{1/(1000Kl),\frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}$, $\eta_y=\eta_x/256$, $\beta=\eta_yK\mu/80000$, $\eta_{x,l}\leq\min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}},\sqrt{\frac{\beta}{6144\eta_xpl^2K^3}},O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, $\eta_{y,l}\leq\min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}},\sqrt{\frac{\eta_y}{3072\eta_xl^2K^2}},O(\epsilon\sqrt{\kappa^{-1}(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, *when* $m=M$ *or* $\sigma_G=0$, *with* $T=\Theta(\kappa\epsilon^{-2}),K=\Theta(\kappa m^{-1}\epsilon^{-2})$, *we can find an* $(\epsilon,\epsilon/\sqrt{\kappa})$*-stationary point of* f *with a per-client sample complexity of* $O(\kappa^2m^{-1}\epsilon^{-4})$ *and a communication complexity of* $O(\kappa\epsilon^{-2})$. *Here,* $\Delta=V_0-\Phi^*$, $\kappa=l/\mu$.

Proof Combining Lemma C.7, Lemma E.3, with $\beta=\eta_yK\mu/80000$, we have

$$\begin{aligned}
\mathbb{E}V_t-\mathbb{E}V_{t+1}\geq&\left(\frac{\eta_xK}{32}-\frac{480\beta}{\mu}-\frac{1920\beta}{l}\right)\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\left(\frac{\eta_yK}{16}-\frac{480\beta}{\mu}\right)\frac{1}{\eta_y^2K^2}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]\\
\geq&\frac{\eta_xK}{64}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{32\eta_yK}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].
\end{aligned}\tag{23}$$

With $T=mK=\Theta(\kappa\epsilon^{-2}),K=\Theta(\kappa m^{-1}\epsilon^{-2})$, $\eta_x=\min\{1/(1000Kl),\frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}=\Theta(\kappa^{-1}m\epsilon^2)$, $\beta=\eta_yK\mu/80000=\Theta(\kappa^{-2}m\epsilon^2)$, when $M=m$ or $\sigma_G=0$, and $\eta_{x,l}^2\leq O(\kappa^{-1}\epsilon^2)K^{-2},\eta_{y,l}^2\leq O(\kappa^{-1}\epsilon^2)K^{-2}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2\leq\frac{O(1)}{\eta_xKT}\Delta+O(1)\frac{\eta_xl\sigma^2}{m}+O(1)\frac{\sigma^2}{mK}+O(1)\kappa^{-1}\epsilon^2\leq O(1)\kappa^{-1}\epsilon^2\tag{24}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_y^2K^2}\mathbb{E}\|\bar{y}_{t+1}-y_t\|^2\leq\frac{O(1)}{\eta_xKT}\Delta+O(1)\frac{\eta_xl\sigma^2}{m}+O(1)\frac{\sigma^2}{mK}+O(1)\kappa^{-1}\epsilon^2\leq O(1)\kappa^{-1}\epsilon^2\tag{25}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} p^2\mathbb{E}\|x_t - z_t\|^2 \leq \frac{O(1)\kappa}{\eta_x KT}\Delta + O(1)\frac{\kappa\eta_x l\sigma^2}{m} + O(1)\frac{\kappa\sigma^2}{mK} + O(1)\epsilon^2 \leq O(1)\epsilon^2 \tag{26}$$

Because $\eta_y K \leq 1/l$, we have

$$\begin{aligned}
&l^2\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \\
\leq &\frac{1}{\eta_y^2 K^2}\|P_Y(y_t + \eta_y K\nabla_y f(x_t, y_t)) - y_t\|^2 \\
= &\frac{1}{\eta_y^2 K^2}\|\bar{y}_{t+1} - y_t\|^2.
\end{aligned}$$

So, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} l^2\mathbb{E}\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \leq O(1)\kappa^{-1}\epsilon^2. \tag{27}$$

According to (12), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 2p^2\mathbb{E}\|x_t - z_t\|^2 \leq O(1)\epsilon^2. \tag{28}$$

Thus, we can find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f, with $K = O(\kappa m^{-1}\epsilon^{-2}), T = O(\kappa\epsilon^{-2})$, which means a per-client sample complexity of $KT = O(\kappa^2 m^{-1}\epsilon^{-4})$ and a communication complexity of $T = O(\kappa\epsilon^{-2})$. ■

Corollary E.1 *Under the Assumptions 2.1, 2.4, 3.2, 3.3, E.1, when $M = 1$, if we apply Algorithm 2 with $p = 2l, \quad \eta_x = 1/(1000Kl), \quad \eta_y = \eta_x/256, \quad \beta = \eta_y K\mu/80000$, we could have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2 + \kappa l^2\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \leq \frac{cl\Delta\kappa}{T},$$

where $\Delta = V_0 - \Phi^*$, $\kappa = l/\mu$, c is an $O(1)$ constant. This implies an sample of $O(\kappa\epsilon^{-2})$ to find an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f.

Proof Applying Algorithm 2 with $p = 2l, \quad \eta_x = 1/(1000l), \quad \eta_y = \eta_x/256, \quad \beta = \eta_y K\mu/80000$ is equivalent to applying Algorithm 1 with $m = M = 1$, $K = 1$, $p = 2l, \quad \eta_x = \min\{1/(1000Kl), \quad \frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}, \quad \eta_y = \eta_x/256, \quad \beta = \eta_y K\mu/80000$ and any appropriate $\eta_{x,l}, \eta_{y,l}$. Thus, according to Theorem E.1 and (23), we have

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq \frac{\eta_x}{64}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{1}{32\eta_y}\mathbb{E}\|\bar{y}_{t+1} - y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 \tag{29}$$

Telescoping and rearranging, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x \hat{f}(w_t, z_t)\|^2 \leq \frac{64}{\eta_x T}\Delta \tag{30}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_y^2}\|\bar{y}_{t+1} - y_t\|^2 \leq \frac{32}{\eta_y T}\Delta \tag{31}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} p^2\|x_t - z_t\|^2 \leq \frac{16}{Tp\beta}\Delta = \frac{O(1)\kappa}{\eta_x T}\Delta \tag{32}$$

Because $\eta_y \leq 1/l$, we have

$$\begin{aligned}
&l^2\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \\
\leq &\frac{1}{\eta_y^2}\|P_Y(y_t + \eta_y\nabla_y f(x_t, y_t)) - y_t\|^2 \\
= &\frac{1}{\eta_y^2}\|\bar{y}_{t+1} - y_t\|^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2 + \kappa l^2\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \\
\leq &\frac{1}{T}\sum_{t=0}^{T-1} 2\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 2p^2\|x_t - z_t\|^2 + \frac{\kappa}{\eta_y^2}\|\bar{y}_{t+1} - y_t\|^2 \\
\leq &\frac{O(1)l\Delta\kappa}{T}.
\end{aligned}$$

■

F Nonconvex-One-Point-Concave

Lemma F.1 *Under the Assumptions 2.1, 3.2, 3.4, we have*

$$\|x^*(z) - x^*(y^+(z), z)\|^2 \leq \frac{2(1 + \eta_y Kl + \eta_y Kl\gamma_2)}{\eta_y K(p - l)}\|y - y^+(z)\|D(Y),$$

where $D(Y)$ is the diameter of Y.

Proof Note that Under the Assumption 3.4, we have

$$\begin{aligned}
&\hat{f}(x^*(y^+(z), z), y^*(x^*(y^+(z), z)), z) - \hat{f}(x^*(y^+(z), z), y^+(z), z) \\
\leq &\langle\nabla_y \hat{f}(x^*(y^+(z), z), y^+(z), z), y^*(x^*(y^+(z), z)) - y^+(z)\rangle \\
= &\langle\nabla_y \hat{f}(x^*(y, z), y, z), y^*(x^*(y^+(z), z)) - y^+(z)\rangle + \\
&\langle\nabla_y \hat{f}(x^*(y^+(z), z), y^+(z), z) - \nabla_y \hat{f}(x^*(y, z), y, z), y^*(x^*(y^+(z), z)) - y^+(z)\rangle \\
\overset{(a)}{\leq} &\frac{1}{\eta_y K}\langle y^+(z) - y, y^*(x^*(y^+(z), z)) - y^+(z)\rangle + \\
&\frac{1}{\eta_y K}\langle y + \eta_y K\nabla_y \hat{f}(x^*(y, z), y, z) - y^+(z), y^*(x^*(y^+(z), z)) - y^+(z)\rangle + \\
&(l + l\gamma_2)\|y - y^+(z)\|\|y^*(x^*(y^+(z), z)) - y^+(z)\| \\
\overset{(b)}{\leq} &\frac{1 + \eta_y Kl + \eta_y Kl\gamma_2}{\eta_y K}\|y - y^+(z)\|\|y^*(x^*(y^+(z), z)) - y^+(z)\| \\
\leq &\frac{1 + \eta_y Kl + \eta_y Kl\gamma_2}{\eta_y K}\|y - y^+(z)\|D(Y),
\end{aligned} \tag{33}$$

where in (a), we use the l-smoothness of f , $\nabla_y \hat{f} = \nabla_y f$ and Lemma B.1, in (b), we use the fact that when Y is a closed, convex set, we have

$$\langle a - P_Y(a), b - P_Y(a)\rangle \leq 0, \quad \forall b \in Y. \tag{34}$$

Then by the strong convexity of $\hat{f}(\cdot, y, z)$, we have

$$\|x^*(y^+(z), z) - x^*(z)\|^2$$

$$
\begin{aligned}
&\leq \frac{2}{p-l}[\hat{f}(x^*(y^+(z),z),\hat{y}^*(z),z) - \hat{f}(x^*(z),\hat{y}^*(z),z)] \\
&\overset{(a)}{\leq} \frac{2}{p-l}[\Phi(x^*(y^+(z),z),z) - \hat{f}(x^*(y^+(z),z),y^+(z),z) + \hat{f}(x^*(y^+(z),z),y^+(z),z) - \Phi(x^*(z),z)] \\
&\overset{(b)}{\leq} \frac{2}{p-l}[\hat{f}(x^*(y^+(z),z),y^*(x^*(y^+(z),z)),z) - \hat{f}(x^*(y^+(z),z),y^+(z),z)] \\
&\overset{(c)}{\leq} \frac{2(1+\eta_y Kl + \eta_y Kl\gamma_2)}{\eta_y K(p-l)}\|y - y^+(z)\|D(Y).
\end{aligned}
$$

(a) can be explained by the fact that $\hat{f}(x^*(y^+(z),z),\hat{y}^*(z),z) \leq \Phi(x^*(y^+(z),z),z)$. (b) arises from the relationship $\hat{f}(x^*(y^+(z),z),y^+(z),z) \leq \Phi(x^*(z),z)$. (c) can be attributed to (33). ■

Lemma F.2 *Under the Assumptions 2.1, 3.2, 3.4, and when $\eta_y \leq 1/(1000Kl), p = 2l$, we have*

$$
\|x^*(z) - x^*(y^+(z),z)\|^2 \leq \frac{6D(Y)}{l\eta_y^2 K^2\epsilon^2}\|y^+(z) - y\|^2 + \frac{2D(Y)}{l}\epsilon^2.
$$

Proof When $\eta_y \leq 1/(1000Kl), p = 2l$, we have

$$
\begin{aligned}
&\|x^*(z) - x^*(y^+(z),z)\|^2 \\
\leq &\frac{2(1+\eta_y Kl + \eta_y Kl\gamma_2)}{\eta_y K(p-l)}\|y - y^+(z)\|D(Y) \\
\leq &\frac{4}{\eta_y Kl}\|y - y^+(z)\|D(Y) \\
\leq &\frac{2D(Y)}{l\eta_y^2 K^2\epsilon^2}\|y^+(z) - y\|^2 + \frac{2D(Y)}{l}\epsilon^2,
\end{aligned}
$$

where in the second inequality, we use Lemma F.1. ■

Lemma F.3 *Under the Assumptions 3.2, 3.3, with $p = 2l, \eta_x \leq 1/(1000Kl), \eta_y = \eta_x/256, \beta \leq \eta_y Kl/80000$, $\eta_{x,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\beta}{6144\eta_x pl^2K^3}}\}, \eta_{y,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\eta_y}{3072\eta_x l^2K^2}}\}$, we have*

$$
\begin{aligned}
&\mathbb{E}V_t - \mathbb{E}V_{t+1} \\
\geq &\frac{\eta_x K}{64}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{1}{64\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - 48p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 - \\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2K\frac{\sigma^2}{m} - 8\eta_y K(M-m)\frac{\sigma_G^2}{mM} - \frac{4\eta_y\sigma^2}{m} - \\
&4\eta_x Kl^2[24K^2(\eta_{x,l}^2 + \eta_{y,l}^2)(\sigma_G^2 + G_y^2) + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2].
\end{aligned}
$$

Proof According to Lemma C.7, we have

$$
\begin{aligned}
&\mathbb{E}V_t - \mathbb{E}V_{t+1} \\
\geq &\frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{1}{16\eta_y K}\mathbb{E}\|\bar{y}_{t+1} - y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - 24p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t,z_t)\|^2 - \\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2K\frac{\sigma^2}{m} - 8\eta_y K(M-m)\frac{\sigma_G^2}{mM} - \frac{4\eta_y\sigma^2}{m} - \\
&4\eta_x Kl^2[24K^2(\eta_{x,l}^2 + \eta_{y,l}^2)(\sigma_G^2 + G_y^2) + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2] \\
\overset{(a)}{\geq} &\frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{1}{32\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 - \frac{1}{8\eta_y K}\mathbb{E}\|\bar{y}_{t+1} - y_t^+(z_t)\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - \\
&48p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t),z_t)\|^2 - 48p\beta\mathbb{E}\|x^*(y_t^+(z_t),z_t) - x^*(y_t,z_t)\|^2 -
\end{aligned}
$$

$$
\begin{aligned}
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]\\
\overset{(b)}{\geq}&\frac{\eta_xK}{32}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{32\eta_yK}\mathbb{E}\|y_t^+(z_t)-y_t\|^2-\frac{l^2}{8\eta_yK}\mathbb{E}\|x^*(y_t,z_t)-x_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-\\
&48p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t^+(z_t),z_t)\|^2-48p\beta\gamma_2^2\mathbb{E}\|y_t^+(z_t)-y_t\|^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]\\
\overset{(c)}{\geq}&\left(\frac{\eta_xK}{32}-\frac{l^2}{8\eta_yK(p-l)^2}\right)\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\left(\frac{1}{32\eta_yK}-48p\beta\gamma_2^2\right)\mathbb{E}\|y_t^+(z_t)-y_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-\\
&48p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t^+(z_t),z_t)\|^2-25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]\\
\geq&\frac{\eta_xK}{64}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{64\eta_yK}\mathbb{E}\|y_t^+(z_t)-y_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-48p\beta\mathbb{E}\|x^*(z_t)-x^*(y_t^+(z_t),z_t)\|^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2],
\end{aligned}
$$

where in (a), we use $\|a\|^2\geq\frac{1}{2}\|a-b\|^2-2\|b\|^2$, in (b), we use Lemma B.1, in (c), we use the $(p-l)$-strongly convexity of $\hat{f}(\cdot,y,z)$. ■

Proof of Theorem 3.2

Theorem 3.2 Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.2, 3.3, 3.4 and $\epsilon^2\leq lD(Y)$, if we apply Algorithm 1 with $p=2l$, $\eta_x=\min\{1/(1000Kl),\frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}$, $\eta_y=\eta_x/256$, $\beta=\eta_yK\epsilon^2/(80000D(Y))$, $\eta_{x,l}\leq\min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}},\sqrt{\frac{\beta}{6144\eta_xpl^2K^3}},O(\epsilon^2\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, $\eta_{y,l}\leq\min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}},\sqrt{\frac{\eta_y}{3072\eta_xl^2K^2}},O(\epsilon^2\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, when $m=M$ or $\sigma_G=0$, with $T=\Theta(\epsilon^{-4}),K=\Theta(m^{-1}\epsilon^{-4})$, we can find an (ϵ,ϵ^2)-stationary point of f and an ϵ-stationary point of $\Phi_{1/2l}$ with a per-client sample complexity of $O(m^{-1}\epsilon^{-8})$ and a communication complexity of $O(\epsilon^{-4})$. Here, $\Delta=V_0-\Phi^*$.

Proof Combining Lemma F.2 and F.3, with $\epsilon^2/D(Y)\leq l$, and $\beta=\eta_yK\epsilon^2/(80000D(Y))\leq\eta_yKl/80000$, we have

$$
\begin{aligned}
\mathbb{E}V_t-\mathbb{E}V_{t+1}\geq&\frac{\eta_xK}{64}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\left(\frac{\eta_yK}{64}-\frac{192\beta D(Y)}{\epsilon^2}\right)\frac{1}{\eta_y^2K^2}\mathbb{E}\|y_t^+(z_t)-y_t\|^2+\\
&\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-192\beta D(Y)\epsilon^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]\\
\geq&\frac{\eta_xK}{64}\mathbb{E}\|\nabla_x\hat{f}(w_t,z_t)\|^2+\frac{1}{128\eta_yK}\mathbb{E}\|y_t^+(z_t)-y_t\|^2+\frac{p\beta}{16}\mathbb{E}\|x_t-z_t\|^2-192\beta D(Y)\epsilon^2-\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM}-15l\eta_x^2K\frac{\sigma^2}{m}-8\eta_yK(M-m)\frac{\sigma_G^2}{mM}-\frac{4\eta_y\sigma^2}{m}-\\
&4\eta_xKl^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)(\sigma_G^2+G_y^2)+3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].
\end{aligned}
\tag{35}
$$

Choosing $T=mK=\Theta(\epsilon^{-4}),K=\Theta(m^{-1}\epsilon^{-4})$, $\eta_x=\min\{1/(1000Kl),\frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}=\Theta(m\epsilon^4)$, when $M=m$ or $\sigma_G=0$, and $\eta_{x,l}^2\leq O(\epsilon^4)K^{-2},\eta_{y,l}^2\leq O(\epsilon^4)K^{-2}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 \leq \frac{O(1)}{\eta_x KT}\Delta + O(1)\frac{\eta_x l\sigma^2}{m} + O(1)\frac{\sigma^2}{mK} + O(1)\epsilon^4 \leq O(1)\epsilon^4, \tag{36}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_y^2K^2}\mathbb{E}\|y_t^+(z_t)-y_t\|^2 \leq \frac{O(1)}{\eta_x KT}\Delta + O(1)\frac{\eta_x l\sigma^2}{m} + O(1)\frac{\sigma^2}{mK} + O(1)\epsilon^4 \leq O(1)\epsilon^4, \tag{37}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}p^2\mathbb{E}\|x_t-z_t\|^2 \leq \frac{O(1)}{\epsilon^2\eta_x KT}\Delta + O(1)\frac{\eta_x l\sigma^2}{m\epsilon^2} + O(1)\frac{\sigma^2}{mK\epsilon^2} + O(1)\epsilon^2 \leq O(1)\epsilon^2. \tag{38}$$

Combining (12), (36), (38), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + 2p^2\mathbb{E}\|x_t-z_t\|^2 \leq O(\epsilon^2). \tag{39}$$

Since $\eta_y K \leq 1/l$, we have

$$\begin{aligned}
&l^2\|P_Y(y_t+1/l\nabla_y f(x_t,y_t))-y_t\|^2\\
\leq&\frac{1}{\eta_y^2K^2}\|P_Y(y_t+\eta_y K\nabla_y f(x_t,y_t))-y_t\|^2\\
\leq&\frac{2}{\eta_y^2K^2}\|P_Y(y_t+\eta_y K\nabla_y \hat{f}(x^*(y_t,z_t),y_t))-y_t\|^2+\\
&\frac{2}{\eta_y^2K^2}\|P_Y(y_t+\eta_y K\nabla_y \hat{f}(x^*(y_t,z_t),y_t))-P_Y(y_t+\eta_y K\nabla_y f(x_t,y_t))\|^2\\
\leq&\frac{2}{\eta_y^2K^2}\|y_t^+(z_t)-y_t\|^2+2l^2\|x^*(y_t,z_t)-x_t\|^2\\
\leq&\frac{2}{\eta_y^2K^2}\|y_t^+(z_t)-y_t\|^2+2l^2/(p-l)^2\|\nabla_x \hat{f}(w_t,z_t)\|^2\\
\leq&\frac{2}{\eta_y^2K^2}\|y_t^+(z_t)-y_t\|^2+2\|\nabla_x \hat{f}(w_t,z_t)\|^2.
\end{aligned} \tag{40}$$

Combining (40),(36), (37), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}l^2\|P_Y(y_t+1/l\nabla_y f(x_t,y_t))-y_t\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\nabla_x \hat{f}(w_t,z_t)\|^2 + \frac{2}{\eta_y^2K^2}\mathbb{E}\|y_t^+(z_t)-y_t\|^2 \leq O(\epsilon^4). \tag{41}$$

According to Lemma B.3, we have

$$\begin{aligned}
&\|\nabla_x \Phi_{1/2l}(x_t)\|^2 = p^2\|x_t-x^*(x_t)\|^2\\
\leq&4p^2\|x_t-x^*(y_t,z_t)\|^2+4p^2\|x^*(y_t,z_t)-x^*(y_t^+(z_t),z_t)\|^2+\\
&4p^2\|x^*(y_t^+(z_t),z_t)-x^*(z_t)\|^2+4p^2\|x^*(z_t)-x^*(x_t)\|^2\\
\leq&\frac{4p^2}{(p-l)^2}\|\nabla_x \hat{f}(w_t,z_t)\|^2+4p^2\gamma_2^2\|y_t-y_t^+(z_t)\|^2+\\
&4p^2\left\{\frac{4D(Y)}{l\eta_y^2K^2\epsilon^2}\|y_t^+(z_t)-y_t\|^2+\frac{2D(Y)}{l}\epsilon^2\right\}+4p^2\gamma_1^2\|z_t-x_t\|^2,
\end{aligned} \tag{42}$$

where in the second inequality, we use the $(p-l)$-strongly convexity of $\hat{f}(\cdot,y,z)$, Lemma B.1 and Lemma F.2. Combining (42), (36), (37), (38), we further have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x \Phi_{1/2l}(x_t)\|^2 \leq O(\epsilon^2). \tag{43}$$

Hence, we can identify an (ϵ, ϵ^2)-stationary point for f and an ϵ-stationary point for $\Phi_{1/2l}$, with respective values of $K = \Theta(m^{-1}\epsilon^{-4})$ and $T = \Theta(\epsilon^{-4})$. This results in a per-client sample complexity of $KT = O(m^{-1}\epsilon^{-8})$ and a communication complexity of $T = O(\epsilon^{-4})$. ■

Proof of Corollary 3.2

Corollary 3.2 Under Assumptions 2.1, 2.4, 3.2, 3.3, 3.4, and when $M = 1$, $\epsilon^2 \leq lD(Y)$, if we apply Algorithm 2 with $p = 2l$, $\eta_x = 1/(1000l)$, $\eta_y = \eta_x/256$, $\beta = \eta_y\epsilon^2/(80000D(Y))$, $T = \Theta(\epsilon^{-4})$, we can find an (ϵ, ϵ^2)-stationary point of f and an ϵ-stationary point of $\Phi_{1/2l}$ with a sample complexity of $O(\epsilon^{-4})$.

Proof Applying Algorithm 2 with $p = 2l$, $\eta_x = 1/(1000l)$, $\eta_y = \eta_x/256$, $\beta = \eta_y\epsilon^2/(80000D(Y))$ is equivalent to applying Algorithm 1 with $p = 2l$, $\eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}$, $\eta_y = \eta_x/256$, $\beta = \eta_y K\epsilon^2/(80000D(Y))$ and any appropriate $\eta_{x,l}, \eta_{y,l}$. Thus, according to Theorem 3.2 and (35), we have

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq \frac{\eta_x}{64}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{1}{128\eta_y}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - 192\beta D(Y)\epsilon^2 \tag{44}$$

Telescoping and rearranging, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x \hat{f}(w_t, z_t)\|^2 \leq \frac{64}{\eta_x T}\Delta + O(1)\epsilon^4 \leq O(\epsilon^4) \tag{45}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_y^2}\|y_t^+(z_t) - y_t\|^2 \leq \frac{128}{\eta_y T}\Delta + O(1)\epsilon^4 \leq O(\epsilon^4) \tag{46}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}p^2\|x_t - z_t\|^2 \leq \frac{16}{Tp\beta}\Delta + O(1)\epsilon^2 \leq O(\epsilon^2) \tag{47}$$

Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x f(x_t, y_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}2\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 2p^2\|x_t - z_t\|^2 \leq O(\epsilon^2). \tag{48}$$

According to (40), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}l^2\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}2\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{2}{\eta_y^2}\|y_t^+(z_t) - y_t\|^2 \leq O(\epsilon^4). \tag{49}$$

With (42), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_x \Phi_{1/2l}(x_t)\|^2 \\
\leq &\frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{4p^2}{(p-l)^2}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 4p^2\gamma_2^2\|y_t - y_t^+(z_t)\|^2 + \right. \\
&\left. 4p^2\left\{\frac{4D(Y)}{l\eta_y^2K^2\epsilon^2}\|y_t^+(z_t) - y_t\|^2 + \frac{2D(Y)}{l}\epsilon^2\right\} + 4p^2\gamma_1^2\|z_t - x_t\|^2\right] \leq O(\epsilon^2)
\end{aligned} \tag{50}$$

Thus, we can identify an (ϵ, ϵ^2)-stationary point for f and an ϵ-stationary point for $\Phi_{1/2l}$ with a sample complexity of $T = O(\epsilon^{-4})$. ■

G Nonconvex-Concave

Proof of Theorem 3.3 and Corollary 3.3

Proof We define $\tilde{f}(x,y) = f(x,y) - \frac{\epsilon}{4D(Y)}\|y-y_0\|^2$. Then $\tilde{f}$ is $(l+\epsilon/2D_Y)$-smooth, and $\epsilon/2D(Y)$-strongly-concave. When $\epsilon \leq 2lD(Y)$, $\tilde{f}$ is $2l$-smooth, we have

$$\kappa' = \frac{2lD(Y)}{\epsilon} = O(\epsilon^{-1}).$$

Proof of Theorem 3.3: According to Theorem E.1, applying Algorithm 1 to optimize $\tilde{f}$, with $M = m$ or $\sigma_G = 0$, we can find $(\hat{x},\hat{y})$, an $(\epsilon, \epsilon/\sqrt{\kappa'})$-stationary point of $\tilde{f}$, with a sample complexity of $O(\kappa'^2 n^{-1}\epsilon^{-4}) = O(n^{-1}\epsilon^{-6})$ and a communication complexity of $O(\kappa'^1\epsilon^{-2}) = O(\epsilon^{-3})$.

$(\hat{x},\hat{y})$ is an $(\epsilon, \epsilon/\sqrt{\kappa'})$-stationary point of $\tilde{f}$ means

$$\|\nabla_x \tilde{f}(\hat{x},\hat{y})\| \leq \epsilon$$
$$\|\nabla_y \tilde{f}(\hat{x},\hat{y})\| \leq \epsilon/\sqrt{\kappa'} \leq \epsilon.$$

By the inequality $\max_{x\in X, y\in Y}\|\nabla f(x,y) - \nabla\tilde{f}(x,y)\| \leq \epsilon/2$, we have

$$\|\nabla f(\hat{x},\hat{y})\| \leq \|\nabla\tilde{f}(\hat{x},\hat{y})\| + \|\nabla f(\hat{x},\hat{y}) - \nabla\tilde{f}(\hat{x},\hat{y})\| \leq 2\epsilon.$$

Therefore, $(\hat{x},\hat{y})$ is a $O(\epsilon)$-stationary point of f. We can find an ϵ-stationary point of f with a per-client sample complexity of $O(n^{-1}\epsilon^{-6})$ and a communication complexity of $O(\epsilon^{-3})$.

Proof of Corollary 3.3: Similarly, according to Corollary E.1, applying Algorithm 2 to optimize $\tilde{f}$, with $M = 1$, we can find $(\hat{x},\hat{y})$, an $(\epsilon, \epsilon/\sqrt{\kappa'})$-stationary point of $\tilde{f}$, with a sample complexity of $O(\kappa'\epsilon^{-2}) = O(\epsilon^{-3})$.

$(\hat{x},\hat{y})$ is an $(\epsilon, \epsilon/\sqrt{\kappa'})$-stationary point of $\tilde{f}$ means

$$\|\nabla_x \tilde{f}(\hat{x},\hat{y})\| \leq \epsilon$$
$$\|\nabla_y \tilde{f}(\hat{x},\hat{y})\| \leq \epsilon/\sqrt{\kappa'} \leq \epsilon.$$

By the inequality $\max_{x\in X, y\in Y}\|\nabla f(x,y) - \nabla\tilde{f}(x,y)\| \leq \epsilon/2$, we have

$$\|\nabla f(\hat{x},\hat{y})\| \leq \|\nabla\tilde{f}(\hat{x},\hat{y})\| + \|\nabla f(\hat{x},\hat{y}) - \nabla\tilde{f}(\hat{x},\hat{y})\| \leq 2\epsilon.$$

Therefore, $(\hat{x},\hat{y})$ is a $O(\epsilon)$-stationary point of f. We can find an ϵ-stationary point of f with a sample complexity of $O(\epsilon^{-3})$. ■

H Minimizing the Point-Wise Maximum of Finite Functions

Lemma H.1 (Lemma B13(Zhang et al., 2020)) *Let $x^+(y,z) = x - \eta_x K\nabla_x\hat{f}(x,y,z)$. If Assumption 3.6 holds for problem (2), then there exists $\delta > 0$, such that if $\|z\|$ is bounded by a constant D_z as*

$$\|z\| \leq D_z,$$

and

$$\max\{\|x - x^+(y,z)\|, \|y - y^+(z)\|, \|x^+(y,z) - z\|\} < \delta,$$

we have

$$\|x(y^+(z),z) - x^*(z)\| < \gamma_3\|y - y^+(z)\|$$

for some constant $\gamma_3 > 0$.

Proof of Theorem 3.4

Theorem 3.2 Under Assumptions 2.1, 2.2, 2.3, 2.4, 3.3, 3.6, if we apply Algorithm 1 with $p = 2l$, $\eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\}$, $\eta_y = \eta_x/256$, $\beta = \min\{\eta_y Kl/80000, \delta/2\lambda_1, \delta/4\lambda_2, \eta_y K\delta/4\lambda_3, \frac{1}{6144p\eta_y K\gamma_3^2}\}$, $\eta_{x,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\beta}{6144\eta_x pl^2K^3}}, O(\epsilon\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, $\eta_{y,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\eta_y}{3072\eta_x l^2K^2}}, O(\epsilon\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, when $m = M$ or $\sigma_G = 0$, with $T = \Theta(\epsilon^{-2}), K = \Theta(m^{-1}\epsilon^{-2})$, we can find an ϵ-stationary point of f and an ϵ-stationary point of $\Phi_{1/2l}$ with a per-client sample complexity of $O(m^{-1}\epsilon^{-4})$ and a communication complexity of $O(\epsilon^{-2})$. Here, $\Delta = V_0 - \Phi^*$, $\delta, \lambda_1, \lambda_2, \lambda_3$ are $O(1)$ constants defined in following proof.

Proof According to Lemma F.3, we have

$$\begin{aligned}
&\mathbb{E}V_t - \mathbb{E}V_{t+1}\\
\geq&\frac{\eta_x K}{64}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{1}{64\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - 48p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 -\\
&25l\eta_x^2K^2(M-m)\frac{\sigma_G^2}{mM} - 15l\eta_x^2K\frac{\sigma^2}{m} - 8\eta_y K(M-m)\frac{\sigma_G^2}{mM} - \frac{4\eta_y\sigma^2}{m} -\\
&4\eta_x Kl^2[24K^2(\eta_{x,l}^2 + \eta_{y,l}^2)(\sigma_G^2 + G_y^2) + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2].
\end{aligned}$$

With $T = mK = \Theta(\epsilon^{-2}), K = \Theta(m^{-1}\epsilon^{-2})$, $\eta_x = \min\{1/(1000Kl), \frac{\sqrt{m\Delta}}{\sigma\sqrt{KTl}}\} = \Theta(m\epsilon^2), \beta \leq \eta_y Kl/80000$, when $M = m$ or $\sigma_G = 0$, and $\eta_{x,l}^2 \leq O(\epsilon^2)K^{-2}, \eta_{y,l}^2 \leq O(\epsilon^2)K^{-2}$, we have

$$\begin{aligned}
&\mathbb{E}V_t - \mathbb{E}V_{t+1}\\
\geq&\frac{\eta_x K}{64}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{1}{64\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 + \frac{p\beta}{16}\mathbb{E}\|x_t - z_t\|^2 - 48p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 -\\
&O(1)\epsilon^2. \qquad (51)
\end{aligned}$$

Note that we assume $\|x_t\| \leq D_x$ for all t, we define $D_z = \max\{D_x, \|z_0\|\}$, then we can prove that, for all t, $\|z_t\| \leq D_z$, we prove it by induction. First for $t = 0$, $\|z_0\| \leq D_z$, we assume when $t = i$, we have $\|z_i\| \leq D_z$, then for $t = i+1$, we have $\|z_{i+1}\| \leq \beta\|x_i\| + (1-\beta)\|z_i\| \leq \beta D_x + (1-\beta)D_z \leq D_z$. So, for all t, we have $\|z_t\| \leq D_z$.

Next, we will prove that for all t, we have

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq \frac{\eta_x K}{128}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{1}{128\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 + \frac{p\beta}{32}\mathbb{E}\|x_t - z_t\|^2 - O(1)\epsilon^2. \qquad (52)$$

For any t, there are two cases.

- **Case 1:**

$$\frac{1}{2}\max\{\frac{\eta_x K}{128}\|\nabla_x \hat{f}(w_t, z_t)\|^2, \frac{1}{128\eta_y K}\|y_t^+(z_t) - y_t\|^2, \frac{p\beta}{32}\|x_t - z_t\|^2\} \leq 48p\beta\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2. \qquad (53)$$

- **Case 2:**

$$\frac{1}{2}\max\{\frac{\eta_x K}{128}\|\nabla_x \hat{f}(w_t, z_t)\|^2, \frac{1}{128\eta_y K}\|y_t^+(z_t) - y_t\|^2, \frac{p\beta}{32}\|x_t - z_t\|^2\} \geq 48p\beta\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2. \qquad (54)$$

For **Case 1**, combining (53) and Lemma F.1, we have

$$\begin{aligned}
&\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\| \leq \frac{2(1 + \eta_y Kl + \eta_y Kl\gamma_2)}{\eta_y K(p-l)}\|y_t - y_t^+(z_t)\|D(Y) \leq \frac{4D(Y)}{\eta_y Kl}\|y_t^+(z_t) - y_t\|\\
&\frac{1}{256\eta_y K}\|y_t^+(z_t) - y_t\|^2 \leq 48p\beta\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 \leq \frac{192p\beta D(Y)}{\eta_y Kl}\|y_t^+(z_t) - y_t\|\\
&\|y_t^+(z_t) - y_t\| \leq O(1)\beta = \lambda_1\beta
\end{aligned}$$

$$\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\| \leq \frac{4D(Y)}{\eta_y K l}\|y_t^+(z_t) - y_t\| \leq O(1)\frac{\beta}{\eta_y K}.$$

This leads to the following results:

$$\begin{aligned}
\|x_t - x_t^+(y_t, z_t)\| =& \eta_x K\|\nabla_x \hat{f}(w_t, z_t)\| \\
\leq& O(1)\sqrt{p\beta\eta_x K}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\| \leq O(1)\frac{\sqrt{\beta\eta_x K}\beta}{\eta_y K} = O(1)\beta = \lambda_2\beta, \\
\|x_t - z_t\| \leq& O(1)\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\| \leq O(1)\frac{\beta}{\eta_y K} = \lambda_3\frac{\beta}{\eta_y K}, \\
\|x_t^+(y_t, z_t) - z_t\| \leq& \eta_x K\|\nabla_x \hat{f}(w_t, z_t)\| + \|x_t - z_t\| \leq \lambda_2\beta + \lambda_3\frac{\beta}{\eta_y K}.
\end{aligned}$$

Therefore, if we choose $\beta = \min\{\eta_y K l/80000, \delta/2\lambda_1, \delta/4\lambda_2, \eta_y K\delta/4\lambda_3\}$, we will have

$$\max\{\|x - x^+(y, z)\|, \|y - y^+(z)\|, \|x^+(y, z) - z\|\} < \delta.$$

According to Lemma H.1, with $\|z_t\| \leq D_z$ and $\beta = \min\{\eta_y K l/80000, \delta/2\lambda_1, \delta/4\lambda_2, \eta_y K\delta/4\lambda_3, \frac{1}{6144p\eta_y K\gamma_3^2}\}$, where $\delta, \lambda_1, \lambda_2, \lambda_3$ are all $O(1)$ constants and are independent of ϵ, we have

$$48p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 \leq 48p\beta\gamma_3^2\mathbb{E}\|y_t^+(z_t) - y_t\|^2 \leq \frac{1}{128\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2. \tag{55}$$

Combining (51) and (55), we get the (52). In **Case 2**, we can easily get the (52). Combining these together, for all t, we have

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \geq \frac{\eta_x K}{128}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{1}{128\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 + \frac{p\beta}{32}\mathbb{E}\|x_t - z_t\|^2 - O(1)\epsilon^2. \tag{56}$$

Note that $\beta = \min\{\eta_y K l/80000, \delta/2\lambda_1, \delta/4\lambda_2, \eta_y K\delta/4\lambda_3, \frac{1}{6144p\eta_y K\gamma_3^2}\}$, where $\delta, \lambda_1, \lambda_2, \lambda_3$ are all $O(1)$ constants and are independent of ϵ, so β is also an $O(1)$ constant and is independent of ϵ, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 \leq \frac{O(1)}{\eta_x K T}\Delta + O(1)\epsilon^2 \leq O(1)\epsilon^2, \tag{57}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\eta_y^2 K^2}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 \leq \frac{O(1)}{\eta_x K T}\Delta + O(1)\epsilon^2 \leq O(1)\epsilon^2, \tag{58}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}p^2\mathbb{E}\|x_t - z_t\|^2 \leq \frac{O(1)}{T}\Delta + O(1)\epsilon^2 \leq O(1)\epsilon^2. \tag{59}$$

Combining (12), (36), (38), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 2p^2\mathbb{E}\|x_t - z_t\|^2 \leq O(\epsilon^2). \tag{60}$$

Combining (40),(57), (58) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}l^2\|P_Y(y_t + 1/l\nabla_y f(x_t, y_t)) - y_t\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + \frac{2}{\eta_y^2 K^2}\mathbb{E}\|y_t^+(z_t) - y_t\|^2 \leq O(\epsilon^2). \tag{61}$$

Note that from previous proof, for any $0 \leq t < T$, we have

$$48p\beta\mathbb{E}\|x^*(z_t) - x^*(y_t^+(z_t), z_t)\|^2 \leq \frac{1}{256\eta_y K}\mathbb{E}\|y_t^+(z_t) - y_t\|^2. \tag{62}$$

According to Lemma B.3, we have

$$\begin{aligned}
&\|\nabla_x \Phi_{1/2l}(x_t)\|^2 = p^2\|x_t - x^*(x_t)\|^2 \\
\leq& 4p^2\|x_t - x^*(y_t, z_t)\|^2 + 4p^2\|x^*(y_t, z_t) - x^*(y_t^+(z_t), z_t)\|^2 + \\
&4p^2\|x^*(y_t^+(z_t), z_t) - x^*(z_t)\|^2 + 4p^2\|x^*(z_t) - x^*(x_t)\|^2 \\
\leq& \frac{4p^2}{(p-l)^2}\|\nabla_x \hat{f}(w_t, z_t)\|^2 + 4p^2\gamma_2^2\|y_t - y_t^+(z_t)\|^2 + \\
&\frac{p}{\eta_y K\beta}\|y_t - y_t^+(z_t)\|^2 + 4p^2\gamma_1^2\|z_t - x_t\|^2,
\end{aligned} \tag{63}$$

where in the second inequality, we use the $(p-l)$-strongly convexity of $\hat{f}(\cdot, y, z)$, Lemma B.1 and (62). Combining (63), (57), (58), (59), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla_x \Phi_{1/2l}(x_t)\|^2 \leq O(\epsilon^2). \tag{64}$$

Therefore, we can find an (ϵ, ϵ^2)-stationary point of f and an ϵ-stationary point of Φ, with $K = \Theta(m^{-1}\epsilon^{-2}), T = \Theta(\epsilon^{-2})$, which means a per-client sample complexity of $KT = O(m^{-1}\epsilon^{-4})$ and a communication complexity of $T = O(\epsilon^{-2})$. ■

I PL-PL

Since $p = 0$ and $Y = \mathbb{R}^{d_2}$ in this section, the updates of `FESS-GDA` are:

$$\begin{aligned}
x_{t+1} &= x_t - \eta_x K(u_{x,t} - e_{x,t}), \\
y_{t+1} &= y_t + \eta_y K(u_{y,t} - e_{y,t}).
\end{aligned}$$

We cite the following known results for ease of exposition.

Lemma I.1 (Nouiehed et al. (2019)) *In the minimax problem, when $-f(x, \cdot)$ satisfies PL condition with constant μ_2 for any x and f satisfies Assumption 2.1, then the function $\Phi(x) := \max_y f(x, y)$ is L-smooth with $L := l + l^2/\mu_2$ and $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$ for any $y^*(x) \in \mathrm{Argmax}_y f(x, y)$.*

Lemma I.2 (Yang et al. (2020)) *In the minimax problem, when the objective function f satisfies Assumption 2.1 (Lipschitz gradient) and the two-sided PL condition with constant μ_1 and μ_2, then function $\Phi(x) := \max_y f(x, y)$ satisfies the PL condition with μ_1.*

Proof of Theorem 3.5

Proof We denote $\kappa_1 = l/\mu_1, \kappa_2 = l/\mu_2$, $\kappa' = \max\{\kappa_1, \kappa_2\}$, $\kappa'' = \min\{\kappa_1, \kappa_2\}$ in this section.

Parameters setting: $p = 0$. When $\mu_1 \geq \mu_2$, we choose $\eta_x = \frac{\eta_y\mu_2^2}{64l^2}$, $\eta_{x,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\eta_x}{1536\eta_y l^2K^2}}, O(\epsilon\kappa'^{-2}\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, $\eta_{y,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{1}{1536l^2K^2}}, O(\epsilon\kappa'^{-2}\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, when $m = M$ or $\sigma_G = 0$, we choose $K = O(1)m^{-1}\kappa'\kappa_1\kappa_2^2\epsilon^{-2}$, $T = O(1)\kappa_1\kappa_2^2\log(\epsilon^{-1}\kappa')$, $\eta_y = \frac{1}{4lK}$, when $m < M$ and $\sigma_G > 0$, we choose $\eta_y = \min\{\frac{1}{4lK}, O(1)m\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}\epsilon^2\}$, $K = O(1)$, $T = O(1)m^{-1}\kappa'\kappa_1^2\kappa_2^4\epsilon^{-2}\log(\epsilon^{-1}\kappa')$.

Conversely, when $\mu_1 \leq \mu_2$, we choose $\eta_y = \frac{\eta_x\mu_1^2}{64l^2}$, $\eta_{x,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{1}{1536l^2K^2}}, O(\epsilon\kappa'^{-2}\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, $\eta_{y,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\eta_y}{1536\eta_x l^2K^2}}, O(\epsilon\kappa'^{-2}\sqrt{(\sigma_G^2+\sigma^2)}(Kl)^{-1})\}$, when $m = M$ or $\sigma_G = 0$, we choose $K = O(1)m^{-1}\kappa'\kappa_2\kappa_1^2\epsilon^{-2}$, $T = O(1)\kappa_2\kappa_1^2\log(\epsilon^{-1}\kappa')$, $\eta_x = \frac{1}{4lK}$, when $m < M$ and $\sigma_G > 0$, we choose $\eta_x = \min\{\frac{1}{4lK}, O(1)m\kappa'^{-1}\kappa_1^{-2}\kappa_2^{-1}\epsilon^2\}$, $K = O(1)$, $T = O(1)m^{-1}\kappa'\kappa_2^2\kappa_1^4\epsilon^{-2}\log(\epsilon^{-1}\kappa')$.

We first consider the proof when $\mu_1 \geq \mu_2$.

Since $\Phi(\cdot)$ is L-smooth, $L = l + \frac{l^2}{\mu_2} = (1+\kappa_2)l$ by Lemma I.1, we have

$$\mathbb{E}\Phi(x_t) - \mathbb{E}\Phi(x_{t+1}) \geq \mathbb{E}\langle \nabla_x \Phi(x_t), x_t - x_{t+1}\rangle - \frac{L}{2}\mathbb{E}\|x_t - x_{t+1}\|^2.$$

Because of the l-smoothness of $f(\cdot)$, we have

$$\begin{aligned}
&\mathbb{E}f(w_{t+1}) - \mathbb{E}f(w_t)\\
\geq&\mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle + \mathbb{E}\langle \nabla_y f(w_t), y_{t+1} - y_t\rangle - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2 - \frac{l}{2}\mathbb{E}\|y_{t+1} - y_t\|^2\\
=&\mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle + \eta_y K\mathbb{E}\langle \nabla_y f(w_t), \nabla_y f(w_t) - \bar{e}_{y,t}\rangle - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2 - \frac{l}{2}\mathbb{E}\|y_{t+1} - y_t\|^2\\
=&\mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle + \eta_y K\mathbb{E}\|\nabla_y f(w_t)\|^2 + \frac{\eta_y K}{2}\mathbb{E}\|\nabla_y f(w_t) - \bar{e}_{y,t}\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 -\\
&\frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2 - \frac{l}{2}\mathbb{E}\|y_{t+1} - y_t\|^2\\
\geq&\mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle + \frac{\eta_y K}{2}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2 - \frac{l}{2}\mathbb{E}\|y_{t+1} - y_t\|^2\\
\overset{(a)}{\geq}&(\frac{\eta_y K}{2} - l\eta_y^2 K^2)\mathbb{E}\|\nabla_y f(w_t)\|^2 - l\eta_y^2 K^2 d_{y,t} - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 + \mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2\\
\overset{(b)}{\geq}&\frac{\eta_y K}{4}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - l\eta_y^2 K^2 d_{y,t} + \mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2,
\end{aligned}$$

where (a) is due to the Lemma B.6, and (b) is due to the condition $\eta_y \leq \frac{1}{4lK}$.

Define $W_t = \Phi(x_t) - \Phi^* + \Phi(x_t) - f(x_t, y_t) = 2\Phi(x_t) - f(x_t, y_t) - \Phi^*$, we have

$$\begin{aligned}
&\mathbb{E}W_t - \mathbb{E}W_{t+1}\\
\geq&2\mathbb{E}\langle \nabla_x \Phi(x_t), x_t - x_{t+1}\rangle - L\mathbb{E}\|x_t - x_{t+1}\|^2 + \frac{\eta_y K}{4}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - l\eta_y^2 K^2 d_{y,t} +\\
&\mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2.
\end{aligned}$$

Denote $A_3 = 2\mathbb{E}\langle \nabla_x \Phi(x_t), x_t - x_{t+1}\rangle - L\mathbb{E}\|x_t - x_{t+1}\|^2 + \mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle - \frac{l}{2}\mathbb{E}\|x_{t+1} - x_t\|^2$, we have

$$\begin{aligned}
A_3 =&2\mathbb{E}\langle \nabla_x \Phi(x_t) - \nabla_x f(w_t), x_t - x_{t+1}\rangle - \frac{2L+l}{2}\mathbb{E}\|x_t - x_{t+1}\|^2 + \mathbb{E}\langle \nabla_x f(w_t), x_{t+1} - x_t\rangle\\
\geq&\eta_y K\mathbb{E}\langle \nabla_x f(w_t), \nabla_x f(w_t) - \bar{e}_{x,t}\rangle - 2\eta_x K\mathbb{E}\|\nabla_x \Phi(x_t) - \nabla_x f(w_t)\|\|\nabla_x f(w_t) - \bar{e}_{x,t}\| - \frac{2L+l}{2}\mathbb{E}\|x_t - x_{t+1}\|^2\\
\overset{(a)}{\geq}&\frac{\eta_x K}{2}\mathbb{E}\|\nabla_x f(w_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - 8\eta_x K\mathbb{E}\|\nabla_x \Phi(x_t) - \nabla_x f(w_t)\|^2 - \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x f(w_t) - \bar{e}_{x,t}\|^2 -\\
&(2L+l)\eta_x^2 K^2\mathbb{E}\|\nabla_x f(w_t)\|^2 - (2L+l)\eta_x^2 K^2 d_{x,t}\\
\overset{(b)}{\geq}&\frac{\eta_x K}{2}\mathbb{E}\|\nabla_x f(w_t)\|^2 - \frac{\eta_x K}{2}\mathbb{E}\|\bar{e}_{x,t}\|^2 - 8\eta_x K l^2\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{\eta_x K}{4}\mathbb{E}\|\nabla_x f(w_t)\|^2 - \frac{\eta_x K}{4}\mathbb{E}\|\bar{e}_{x,t}\|^2 -\\
&(2L+l)\eta_x^2 K^2\mathbb{E}\|\nabla_x f(w_t)\|^2 - (2L+l)\eta_x^2 K^2 d_{x,t}\\
\overset{(c)}{\geq}&\left(\frac{\eta_x K}{4} - \eta_x^2 K^2(2L+l)\right)\mathbb{E}\|\nabla_x f(w_t)\|^2 - \frac{3\eta_x K}{4}\mathbb{E}\|\bar{e}_{x,t}\|^2 - (2L+l)\eta_x^2 K^2 d_{x,t} - \frac{8\eta_x K l^2}{\mu_2^2}\mathbb{E}\|\nabla_y f(w_t)\|^2\\
\overset{(d)}{\geq}&\frac{\eta_x K}{8}\mathbb{E}\|\nabla_x f(w_t)\|^2 - \frac{8\eta_x K l^2}{\mu_2^2}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{3\eta_x K}{4}\mathbb{E}\|\bar{e}_{x,t}\|^2 - (2L+l)\eta_x^2 K^2 d_{x,t},
\end{aligned}$$

where (a) is due to Lemma B.6, (b) is due to l-smoothness of f, (c) is due to μ_2-PL condition of $f(x,\cdot)$, (d) is due to the condition $\eta_x = \frac{\eta_y}{64\kappa_2^2} \leq \frac{1}{8(2L+l)K}$.

Then, we have

$$\mathbb{E}W_t - \mathbb{E}W_{t+1}$$

$$
\begin{aligned}
&\geq \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x f(w_t)\|^2 + \left(\frac{\eta_y K}{4} - \frac{8\eta_x K l^2}{\mu_2^2}\right)\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - l\eta_y^2 K^2 d_{y,t} - \\
&\quad \frac{3\eta_x K}{4}\mathbb{E}\|\bar{e}_{x,t}\|^2 - (2L+l)\eta_x^2 K^2 d_{x,t} \\
&\overset{(a)}{\geq} \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x f(w_t)\|^2 + \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t)\|^2 - \frac{\eta_y K}{2}\mathbb{E}\|\bar{e}_{y,t}\|^2 - l\eta_y^2 K^2 d_{y,t} - \\
&\quad \frac{3\eta_x K}{4}\mathbb{E}\|\bar{e}_{x,t}\|^2 - (2L+l)\eta_x^2 K^2 d_{x,t} \\
&\overset{(b)}{\geq} \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x f(w_t)\|^2 + \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t)\|^2 - (\eta_y K + 2l\eta_y^2 K^2)\mathbb{E}\|\bar{e}_{y,t}\|^2 - 2l\eta_y^2 K\frac{\sigma^2}{m} - 4l\eta_y^2 K^2(M-m)\frac{\sigma_G^2}{mM} - \\
&\quad (\eta_x K + 2(2L+l)\eta_x^2 K^2)\mathbb{E}\|\bar{e}_{x,t}\|^2 - 2(2L+l)\eta_x^2 K\frac{\sigma^2}{m} - 4(2L+l)\eta_x^2 K^2(M-m)\frac{\sigma_G^2}{mM} \\
&\overset{(c)}{\geq} \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x f(w_t)\|^2 + \frac{\eta_y K}{8}\mathbb{E}\|\nabla_y f(w_t)\|^2 - 4l\eta_y^2 K\frac{\sigma^2}{m} - 8l\eta_y^2 K^2(M-m)\frac{\sigma_G^2}{mM} - \\
&\quad 4\eta_y K l^2[24K^2\eta_{x,l}^2\mathbb{E}\|\nabla_x f(w_t)\|^2 + 24K^2\eta_{y,l}^2\mathbb{E}\|\nabla_y f(w_t)\|^2 + 24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2] \\
&\overset{(d)}{\geq} \frac{\eta_x K}{16}\mathbb{E}\|\nabla_x f(w_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 - 4l\eta_y^2 K\frac{\sigma^2}{m} - 8l\eta_y^2 K^2(M-m)\frac{\sigma_G^2}{mM} - \\
&\quad 4\eta_y K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2],
\end{aligned}
$$

where (a) is due to the condition $\eta_x = \frac{\eta_y\mu_2^2}{64l^2} = \frac{\eta_y}{64\kappa_2^2}$, (b) is due to Lemma B.5, (c) is due to Lemma B.4, (d) is due to the condition $\eta_{x,l}^2 \leq \frac{\eta_x}{1536K^2\eta_y l^2}, \eta_{y,l} \leq \frac{1}{1536K^2l^2}$.

Note that

$$
\begin{aligned}
&\frac{\eta_x K}{16}\mathbb{E}\|\nabla_x f(w_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 \\
&\geq \frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \Phi(x_t)\|^2 - \frac{\eta_x K}{8}\mathbb{E}\|\nabla_x \Phi(x_t) - \nabla_x f(w_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 \\
&\geq \frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \Phi(x_t)\|^2 - \frac{\eta_x K l^2}{8}\mathbb{E}\|y_t - y^*(x_t)\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 \\
&\overset{(a)}{\geq} \frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \Phi(x_t)\|^2 - \frac{\eta_x K l^2}{8\mu_2^2}\mathbb{E}\|\nabla_y f(w_t\|^2 + \frac{\eta_y K}{16}\mathbb{E}\|\nabla_y f(w_t)\|^2 \\
&\overset{(b)}{\geq} \frac{\eta_x K}{32}\mathbb{E}\|\nabla_x \Phi(x_t)\|^2 + \frac{\eta_y K}{32}\mathbb{E}\|\nabla_y f(w_t)\|^2 \\
&\overset{(c)}{\geq} \frac{\eta_x K\mu_1}{16}(\Phi(x_t) - \Phi^*) + \frac{\eta_y K\mu_2}{16}\mathbb{E}(\Phi(x_t) - f(x_t, y_t)) \\
&\geq \frac{\eta_x K\mu_1}{16}\mathbb{E}(\Phi(x_t) - \Phi^* + \Phi(x_t) - f(x_t, y_t)) \\
&\geq \frac{\eta_x K\mu_1}{16}\mathbb{E}W_t,
\end{aligned}
$$

where (a) is due to μ_2-PL condition of $f(x,\cdot)$, (b) is due to the condition $\eta_x = \frac{\eta_y\mu_2^2}{64l^2} = \frac{\eta_y}{64\kappa_2^2}$, and (c) is due to the two-side PL condition of f and Lemma I.2.

Thus, we have

$$
\begin{aligned}
\mathbb{E}W_t - \mathbb{E}W_{t+1} \geq &\frac{\eta_x K\mu_1}{16}\mathbb{E}W_t - 4l\eta_y^2 K\frac{\sigma^2}{m} - 8l\eta_y^2 K^2(M-m)\frac{\sigma_G^2}{mM} - \\
&4\eta_y K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2].
\end{aligned}
$$

By telescoping and rearranging, we have

$$
\begin{aligned}
\mathbb{E}W_{t+1} \leq &\left(1 - \frac{\eta_x K\mu_1}{16}\right)\mathbb{E}W_t - 4l\eta_y^2 K\frac{\sigma^2}{m} - 8l\eta_y^2 K^2(M-m)\frac{\sigma_G^2}{mM} - \\
&4\eta_y K l^2[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2 + 2K\eta_{y,l}^2)\sigma^2],
\end{aligned}
$$

$$\begin{aligned}\mathbb{E}W_t \leq & \left(1-\frac{\eta_x K\mu_1}{16}\right)^t \mathbb{E}W_0 + \frac{64l\eta_y^2\sigma^2}{\eta_x\mu_1 m} + \frac{126l\eta_y^2 K}{\eta_x\mu_1}(M-m)\frac{\sigma_G^2}{mM} + \\ & \frac{64\eta_y l^2}{\eta_x\mu_1}[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2] \\ = & \left(1-\frac{\eta_y Kl}{256\kappa_1\kappa_2^2}\right)^t \mathbb{E}W_0 + O(1)\frac{\kappa_1\kappa_2^2\eta_y\sigma^2}{m} + O(1)\kappa_1\kappa_2^2\eta_y K(M-m)\frac{\sigma_G^2}{mM} + \\ & O(1)\kappa_1\kappa_2^2 l[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2].\end{aligned}$$

Note that

$$\mathbb{E}\|x_t - x^*\|^2 \leq \frac{2}{\mu_1}\mathbb{E}(\Phi(x_t)-\Phi^*),$$

$$\mathbb{E}\|y_t - y^*\|^2 \leq 2\mathbb{E}\|y_t - y^*(x_t)\|^2 + 2\mathbb{E}\|y^*(x_t) - y^*\|^2 \leq \frac{2}{\mu_2}(\Phi(x_t) - f(x_t,y_t)) + \frac{2}{\mu_1}\mathbb{E}(\Phi(x_t)-\Phi^*).$$

$$\begin{aligned}\mathbb{E}\|x_t - x^*\|^2 + \mathbb{E}\|y_t - y^*\|^2 \leq & O(1)\kappa'\left(1-\frac{\eta_y Kl}{256\kappa_1\kappa_2^2}\right)^t \mathbb{E}W_0 + O(1)\frac{\kappa'\kappa_1\kappa_2^2\eta_y\sigma^2}{m} + O(1)\kappa'\kappa_1\kappa_2^2\eta_y K(M-m)\frac{\sigma_G^2}{mM} + \\ & O(1)\kappa'\kappa_1\kappa_2^2 l[24K^2(\eta_{x,l}^2+\eta_{y,l}^2)\sigma_G^2 + 3K(\eta_{x,l}^2+2K\eta_{y,l}^2)\sigma^2]. \qquad (65)\end{aligned}$$

When $M = m$ or $\sigma_G = 0$, with $K = O(1)m^{-1}\kappa'\kappa_1\kappa_2^2\epsilon^{-2}$, $\eta_y = 1/(4lK) = O(1)m\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}\epsilon^2$, $T = O(1)\kappa_1\kappa_2^2\log(\epsilon^{-1}\kappa')$, $\eta_{x,l}^2 \leq O(1)\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}K^{-2}\epsilon^2$, $\eta_{x,l}^2 \leq O(1)\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}K^{-2}\epsilon^2$, we have

$$\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq O(1)\epsilon^2,$$

which means a per-client sample complexity of $O(m^{-1}\kappa'\kappa_1^2\kappa_2^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$, a communication complexity of $O(\kappa_1\kappa_2^2\log(\epsilon^{-1}\kappa'))$.

When $m < M$ and $\sigma_G > 0$, with $\eta_y = O(1)m\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}\epsilon^2$, $K = O(1)$, $T = O(1)m^{-1}\kappa'\kappa_1^2\kappa_2^4\epsilon^{-2}\log(\epsilon^{-1}\kappa')$, $\eta_{x,l}^2 \leq O(1)\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}K^{-2}\epsilon^2$, $\eta_{x,l}^2 \leq O(1)\kappa'^{-1}\kappa_1^{-1}\kappa_2^{-2}K^{-2}\epsilon^2$, we have

$$\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq O(1)\epsilon^2,$$

which means both per-client sample complexity and communication complexity are $O(m^{-1}\kappa'\kappa_1^2\kappa_2^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$. Using Kakutoni's Theorem, we have

$$\min_x\max_y f(x,y) = \max_y\min_x f(x,y) = \min_y\max_x(-f(x,y)) = \min_y\max_x g(y,x),$$

where we denote $g(y,x) = -f(x,y)$.

Thus, the minimax problem of a function with μ_1-PL-μ_2-PL is equivalent to minimax problem of a function with μ_2-PL-μ_1-PL. When $M = m$ or $\sigma_G = 0$, it is guaranteed to find x_T, y_T satisfying $\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq O(1)\epsilon^2$ with a per-client sample complexity of $O(m^{-1}\kappa'\kappa_1^4\kappa_2^2\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$ and a communication complexity of $O(\kappa_1^2\kappa_2\log(\epsilon^{-1}\kappa'))$.

Overall, when $M = m$ or $\sigma_G = 0$, we can find x_T, y_T satisfying $\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq O(1)\epsilon^2$ with a per-client sample complexity of $O(m^{-1}\kappa'^3\kappa''^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$ and a communication complexity of $O(\kappa'\kappa''^2\log(\epsilon^{-1}\kappa'))$, where $\kappa' = \max\{\kappa_1,\kappa_2\}, \kappa'' = \min\{\kappa_1,\kappa_2\}$.

Similarly, when $m < M$ and $\sigma_G > 0$, we we can find x_T, y_T satisfying $\mathbb{E}\|x_T - x^*\|^2 + \mathbb{E}\|y_T - y^*\|^2 \leq O(1)\epsilon^2$ with a per-client sample complexity of $O(m^{-1}\kappa'^3\kappa''^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$ and a communication complexity of $O(m^{-1}\kappa'^3\kappa''^4\epsilon^{-2}\log(\epsilon^{-1}\kappa'))$, where $\kappa' = \max\{\kappa_1,\kappa_2\}, \kappa'' = \min\{\kappa_1,\kappa_2\}$. ■

J Proof of Proposition 3.1

Proof According to Proposition 2.1 and (7) in Yang et al. (2022b), if $(\tilde{x}, \tilde{y})$ is an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f, then $\|\nabla_x \Phi_{1/2l}(\tilde{x})\| \leq 2\sqrt{2}\epsilon$. If we could find $\hat{x}$ such that $\mathbb{E}\|\hat{x} - x^*(\tilde{x})\| \leq \frac{\epsilon}{\kappa l}$, then

$$\begin{aligned}\mathbb{E}\|\nabla_x \Phi(\hat{x})\| &\leq \mathbb{E}\|\nabla_x \Phi(x^*(\tilde{x}))\| + \mathbb{E}\|\nabla_x \Phi(\hat{x}) - \nabla_x \Phi(x^*(\tilde{x}))\| \\ &\leq \mathbb{E}\|\nabla_x \Phi_{1/2l}(\tilde{x})\| + 2\kappa l \mathbb{E}\|\hat{x} - x^*(\tilde{x})\| \\ &\leq (2\sqrt{2} + 2)\epsilon,\end{aligned}$$

where the second inequality is because of Lemma B.3 and Lemma I.1. Note that $x^*(\tilde{x})$ is the solution to $\min_x \max_y \tilde{f}(x, y) = \min_x \max_y f(x, y) + l\|x - \tilde{x}\|^2$.

Note that $\tilde{f}(x, y) = f(x, y) + l\|x - \tilde{x}\|^2$ is $3l$-smooth, l-strongly convex in x, μ-PL in y. According to Theorem 3.5, we can use `FESS-GDA` to optimize $\tilde{f}(x, y)$ from initial point $(\tilde{x}, \tilde{y})$. Furthermore, according to (65), with $\eta_x \leq 1/(4lK), \eta_y = \frac{\eta_x \mu_1^2}{64l^2}$, $\eta_{x,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{1}{1536l^2K^2}}, O(\epsilon\kappa^{-3}\sqrt{(\sigma_G^2 + \sigma^2)}(Kl)^{-1})\}$, $\eta_{y,l} \leq \min\{\frac{1}{2l\sqrt{2(2K-1)(K-1)}}, \sqrt{\frac{\eta_y}{1536\eta_x l^2 K^2}}, O(\epsilon\kappa^{-3}\sqrt{(\sigma_G^2 + \sigma^2)}(Kl)^{-1})\}$, we have

$$\mathbb{E}\|x_t - x^*\|^2 \leq O(1)\kappa\left(1 - \frac{\eta_x K l}{256 \times 9\kappa}\right)^t \mathbb{E}W_0 + O(1)\frac{\kappa^2 \eta_x \sigma^2}{m} + O(1)\kappa^2 \eta_x K(M - m)\frac{\sigma_G^2}{mM} + O(1)\kappa^{-2}\epsilon^2,$$

where since $\min_x \max_y \tilde{f}(x, y) = \min_y \max_x(-\tilde{f}(x, y))$, we redefine $W = \tilde{\Psi}^* - \tilde{\Psi}(y) + \tilde{f}(x, y) - \tilde{\Psi}(y)$, $\tilde{\Psi}(y) = \min_x \tilde{f}(x, y) = \min_x(-\tilde{f}(x, y))$. We then have

$$\begin{aligned}W_0 =& \tilde{\Psi}^* - \tilde{\Psi}(\tilde{y}) + \tilde{f}(\tilde{x}, \tilde{y}) - \tilde{\Psi}(\tilde{y}) \\ \overset{(a)}{\leq}& \tilde{\Psi}^* - \tilde{\Psi}(\tilde{y}) + \frac{1}{2l}\|\nabla_x \tilde{f}(\tilde{x}, \tilde{y})\|^2 \\ =& \max_y \min_x \tilde{f}(x, y) - \max_y \tilde{f}(\tilde{x}, y) + \max_y \tilde{f}(\tilde{x}, y) - \tilde{f}(\tilde{x}, \tilde{y}) + \tilde{f}(\tilde{x}, \tilde{y}) - \min_x \tilde{f}(x, \tilde{y}) + \frac{1}{2l}\|\nabla_x \tilde{f}(\tilde{x}, \tilde{y})\|^2 \\ \overset{(b)}{\leq}& \frac{1}{\mu}\|\nabla_y \tilde{f}(\tilde{x}, \tilde{y})\|^2 + \frac{1}{l}\|\nabla_x \tilde{f}(\tilde{x}, \tilde{y})\|^2 + \frac{1}{2l}\|\nabla_x \tilde{f}(\tilde{x}, \tilde{y})\|^2 \\ \overset{(c)}{\leq}& \frac{2\epsilon^2}{l},\end{aligned}$$

where (a) is due to l-strongly convexness of x, (b) is due to l-strongly convexness of x and μ-PL of y, (c) is because that $(\tilde{x}, \tilde{y})$ is an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point of f. Thus, we have

$$\mathbb{E}\|x_t - x^*\|^2 \leq O(1)\kappa\epsilon^2\left(1 - \frac{\eta_x K l}{256 \times 9\kappa}\right)^t + O(1)\frac{\kappa^2 \eta_x \sigma^2}{m} + O(1)\kappa^2 \eta_x K(M - m)\frac{\sigma_G^2}{mM} + O(1)\kappa^{-2}\epsilon^2,$$

Therefore, when $m = M$ or $\sigma_G = 0$, with $\eta_x = 1/(4lK) = O(1)m\epsilon^2\kappa^{-4}$, $K = O(1)m^{-1}\epsilon^{-2}\kappa^4$, $T = O(1)\kappa\log(\kappa)$ we can find $\hat{x}$ such that $\mathbb{E}\|\hat{x} - x^*(\tilde{x})\| \leq O(\frac{\epsilon}{\kappa})$ and $\mathbb{E}\|\nabla_x \Phi(\hat{x})\| \leq O(\epsilon)$ with $KT = O(m^{-1}\kappa^5\epsilon^{-2}\log(\kappa))$ per-client sample complexity and $T = O(\kappa\log(\kappa))$ communication complexity. When $m < M$ and $\sigma_G > 0$, with $K = O(1)$, $\eta_x = \min\{1/(4lK), O(1)m\epsilon^2\kappa^{-4}\} = O(1)m\epsilon^2\kappa^{-4}$, $T = O(1)m^{-1}\kappa^5\epsilon^{-2}\log(\kappa)$, we can find $\hat{x}$ such that $\mathbb{E}\|\hat{x} - x^*(\tilde{x})\| \leq O(\frac{\epsilon}{\kappa})$ and $\mathbb{E}\|\nabla_x \Phi(\hat{x})\| \leq O(\epsilon)$ with $KT = O(m^{-1}\kappa^5\epsilon^{-2}\log(\kappa))$ per-client sample complexity and $T = O(m^{-1}\kappa^5\epsilon^{-2}\log(\kappa))$ communication complexity. ■

K Additional Experiments

Fair Classification

For the fair classification task, we have presented the average test accuracy results in Section 4.2. To compare the fairness of models trained with `FESS-GDA` and Fed-Norm-SGDA+, following the same setting in Section 4.2,

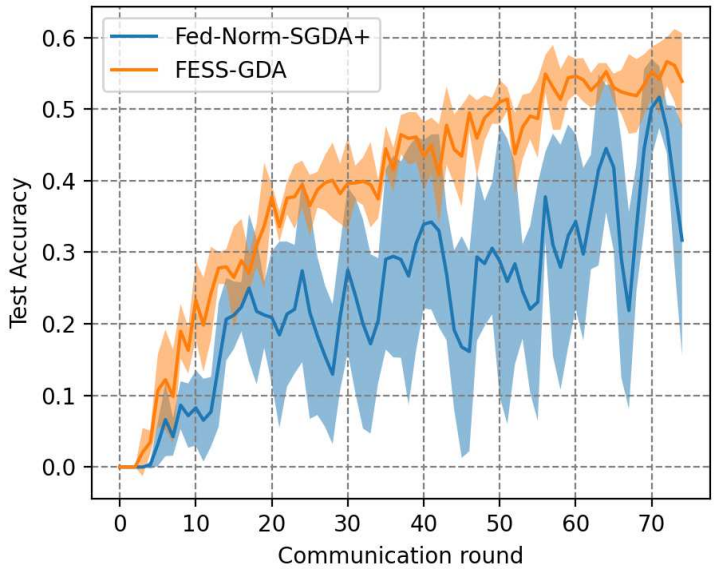

Figure 3: Comparison between Fed-Norm-SGDA+ and `FESS-GDA` for the worst test accuracy over 10 categories of CIFAR-10.

we now present the worst-case test accuracy of models over 10 categories in Figure 3. Figure 2 and Figure 3 show that models trained with `FESS-GDA` not only have better average test accuracy over all categories, but also have better worst-case test accuracy over all categories, which demonstrates that models trained with `FESS-GDA` have better overall performance as well as fairness compared to models trained with Fed-Norm-SGDA+.

Communication Savings from Multiple Local Updates

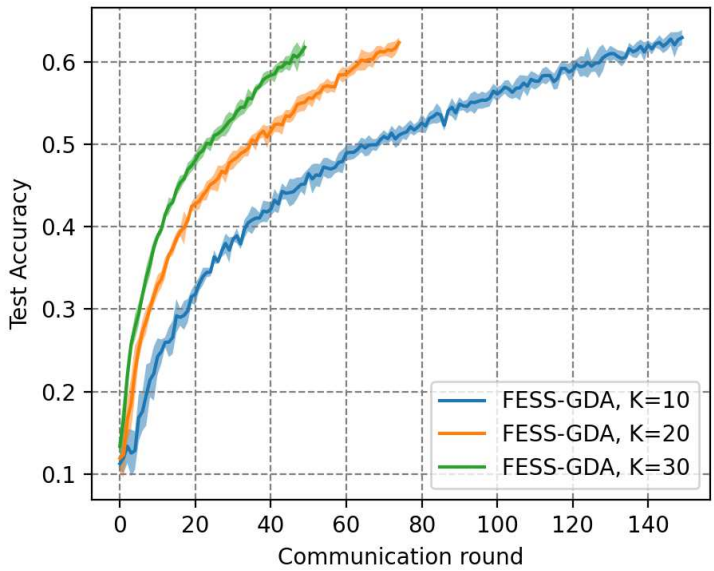

Figure 4: `FESS-GDA` for the fair classification task on CIFAR-10 with different number of local updates.

We test `FESS-GDA` for the fair classification task on the CIFAR10 dataset using the same setting as in Section 4.2

with $\eta_{x,l} = \eta_{y,l} = 0.1, \eta_{x,g} = \eta_{y,g} = 1, p = 0.1, \beta = 0.9$ and number of local updates K from $\{10, 20, 30\}$. Each experiment is repeated 5 times and we report the average performance. As we can see from Figure 4, `FESS-GDA` has significant communication savings from multiple local updates.

Model Architecture for Fair Classification

Table 3 shows the architecture of the convolutional neural network we used for the fair classification task.

Table 3: Model Architecture for CIFAR10 dataset

Layer Type	Shape	padding
Convolution + ReLU	$3 \times 3 \times 16$	1
Max Pooling	2×2	
Convolution + ReLU	$3 \times 3 \times 32$	1
Max Pooling	2×2	
Convolution + ReLU	$3 \times 3 \times 64$	1
Max Pooling	2×2	
Fully Connected + ReLU	512	
Fully Connected + ReLU	64	
Fully Connected	10	