

---

# Multiplayer Federated Learning: Reaching Equilibrium with Less Communication

---

TaeHo Yoon    Sayantan Choudhury    Nicolas Loizou

Department of Applied Mathematics & Statistics

Mathematical Institute for Data Science

Johns Hopkins University

{tyoon7, schoudh8, nloizou}@jhu.edu

## Abstract

Traditional Federated Learning (FL) approaches assume collaborative clients with aligned objectives working toward a shared global model. However, in many real-world scenarios, clients act as rational players with individual objectives and strategic behavior, a concept that existing FL frameworks are not equipped to adequately address. To bridge this gap, we introduce *Multiplayer Federated Learning (MpFL)*, a novel framework that models the clients in the FL environment as players in a game-theoretic context, aiming to reach an equilibrium. In this scenario, each player tries to optimize their own utility function, which may not align with the collective goal. Within MpFL, we propose *Per-Player Local Stochastic Gradient Descent (PEARL-SGD)*, an algorithm in which each player/client performs local updates independently and periodically communicates with other players. We theoretically analyze **PEARL-SGD** and prove that it reaches a neighborhood of equilibrium with less communication in the stochastic setting compared to its non-local counterpart. Finally, we verify our theory through numerical experiments.

## 1 Introduction

Federated Learning (FL) has emerged as a powerful collaborative learning paradigm where multiple clients jointly train a machine learning model without sharing their local data. In the classical FL setting, a central server coordinates multiple clients (e.g., mobile or edge devices) to collaboratively learn a shared global model without exchanging their own training data [59, 66, 94, 77]. In this scenario, each client performs local computations on its private data and periodically communicates model updates to the server, which aggregates them to update the global model. This collaborative approach has been successfully applied in various domains, including natural language processing [83, 52], computer vision [84, 76], and healthcare [5, 135].

Despite their success, traditional FL frameworks rely on the key assumption that all participants are fully cooperative and share aligned objectives, collectively working to optimize the performance of a shared global model (e.g., minimizing the average of individual loss functions). This assumption overlooks situations where participants have individual objectives or competitive interests that may not align with the collective goal. A variety of such scenarios have been extensively considered in the game theory literature, including Cournot competition in economics [2], optical networks [108], electricity markets [116], energy consumption control in smart grid [141], or mobile robot control [60]. These applications have yet to be associated with FL, presenting an underexplored opportunity to bridge game theory and FL for more robust and realistic frameworks.

To address these limitations of classical FL approaches, we propose a novel framework, *Multiplayer Federated Learning (MpFL)*, which models the FL process as a game among rational players with individual utility functions. In MpFL, each participant is considered a player who aims to optimize

their own objective while interacting strategically with other players in the network via a central server. This game-theoretic perspective acknowledges that participants may act in their self-interest, have conflicting goals, or be unwilling to fully cooperate. By incorporating these dynamics, MpFL provides a more realistic and flexible foundation for FL in competitive and heterogeneous environments.

In the literature, there are multiple strategies that aim to incorporate personalization into classical FL, including multi-task learning [124, 98], transfer learning [64], and mixing of the local and global models [49, 50], to name a few. However, to the best of our knowledge, none of them can formulate the behavior of the clients/players in a non-cooperative environment. This gap is precisely what Multiplayer Federated Learning (MpFL) aims to address.

### 1.1 Main contributions

- **Introducing Multiplayer Federated Learning.** We develop a novel framework of *Multiplayer Federated Learning (MpFL)*, which models the FL process as a game among rational players with individual utility functions. In MpFL, each client within the FL environment is viewed as a player of the game, and their local models are viewed as their actions. Each player constantly adjusts their model (action) to optimize their own objective function, and the MpFL framework aims for each player to reach to a Nash equilibrium by collaboratively training their model under the orchestration of a central server (e.g., service provider), while keeping the training data decentralized. That is, MpFL extends the scope of FL to scenarios where clients are allowed to have more general, diversified, possibly competing objectives.
- **Design and analysis of Per-Player Local SGD.** To handle the Multiplayer Federated Learning framework, we introduce *Per-Player Local SGD* (**PEARL-SGD**), a new algorithm inspired by the stochastic gradient descent ascent method in minimax optimization, that is able to handle the competitive nature of the players/clients. In **PEARL-SGD**, each player performs local SGD steps independently on their own actions/strategies (keeping the strategies of the other players fixed), and the updated actions/models are periodically communicated with the other players of the network via a central server.
- **Convergence guarantees for PEARL-SGD on heterogeneous data.** We provide tight convergence guarantees for **PEARL-SGD**, in both deterministic and stochastic regimes with heterogeneous data (see Table 1 for a summary of our results).
  - **Deterministic setting:** For the full-batch (deterministic) variant of **PEARL-SGD**, we prove that under suitable assumptions, **PEARL-SGD** converges linearly to an equilibrium for any communication period  $\tau > 1$ , provided that the constant step-size  $\gamma$  is sufficiently small (see Theorem 3.3).
  - **Stochastic setting:** In its more general version, **PEARL-SGD** assumes that each player uses an unbiased estimator of its gradient in the update rule. For this setting, we provide two Theorems based on two different step-size choices:
    - \* *Constant step-size:* We show that under the same assumptions as in the deterministic case, **PEARL-SGD** converges linearly to a neighborhood of equilibrium (see Theorem 3.4). In Corollary 3.5, we show that with appropriate step-size depending on the total number of local SGD iterations  $T$ , **PEARL-SGD** achieves  $\tilde{O}(1/T)$  convergence rate with improved communication complexity when  $T$  is sufficiently large.
    - \* *Decreasing step-size rule:* We prove that **PEARL-SGD** converges to an exact equilibrium (without neighborhood of convergence) with sublinear convergence (see Theorem 3.6). In this scenario, the asymptotic rate and communication complexity are essentially the same as in Corollary 3.5, but this result does not require the step-sizes to depend on  $T$ .
- **Numerical Evaluation:** We provide numerical experiments verifying our theoretical results and show the benefits in terms of communications of **PEARL-SGD** over its non-local counterpart in the MpFL settings.

## 2 Multiplayer Federated Learning: Definition and Related Work

In this section, we introduce the Multiplayer Federated Learning (MpFL) framework and explain its main differences compared to the classical FL [59], federated minimax optimization [25, 122, 145] and personalized FL [34, 129].

Table 1: Summary of theoretical results for **PEARL-SGD**. Theorem 3.3 considers the full-batch (deterministic) scenario. Theorem 3.4 and Theorem 3.6 both consider the general stochastic case. These results differ in the step-size choice; the former uses a constant step-size, while the latter uses decreasing step-sizes. In the *Convergence* column, “Linear” and “Sublinear” indicates the convergence rate, “Exact” refers to convergence to an equilibrium, and “Neighborhood” refers to convergence to a neighborhood of an equilibrium.

Theorem	Setting	Step-size	Convergence
Theorem 3.3	Deterministic	Constant	Linear+Exact
Theorem 3.4	Stochastic	Constant	Linear+Neighborhood
Theorem 3.6	Stochastic	Decreasing	Sublinear+Exact

## 2.1 Definition of MpFL

*Multiplayer Federated Learning (MpFL)* is a machine learning setting that combines the benefits of a game-theoretic formulation with classical federated learning. In this setting, the problem is an  $n$ -player game in which multiple players/clients (e.g. mobile devices or whole organizations) communicate with each other via a central server (e.g. service provider) to reach equilibrium. That is, reach a set of strategies where no player can unilaterally deviate from their strategy to achieve a better payoff, given the strategies chosen by all other players.

In classical  $n$ -player games, communication between players was assumed to be cheap, easy, and straightforward, mainly because all players were in close proximity and had direct access to one another. This assumption made communication an insignificant concern in typical game theory analysis. However, with the advent of new large-scale machine learning applications, this is no longer the case. Communication between players can be expensive and challenging, especially in distributed systems where the clients/players are geographically dispersed or operate under communication constraints. Addressing communication costs and designing communication-efficient algorithms for  $n$ -player games have become increasingly important, and this is precisely the challenge that Multiplayer Federated Learning aims to address.

**Equilibrium in  $n$ -player game.** Let  $x^i \in \mathbb{R}^{d_i}$  denote the action of player  $i = 1, \dots, n$  and let  $\mathbf{x} = (x^1, \dots, x^n) \in \mathbb{R}^D = \mathbb{R}^{d_1 + \dots + d_n}$  be the joint action/strategy vector of all players. Let  $f_i(x^1, \dots, x^n): \mathbb{R}^{d_1 + \dots + d_n} \rightarrow \mathbb{R}$  be the function of the player  $i$  (which player  $i$  prefers to minimize in  $x^i$ ) and let  $x^{-i} = (x^1, \dots, x^{i-1}, x^{i+1}, \dots, x^n) \in \mathbb{R}^{D-d_i}$  be the vector containing all players' actions except that of player  $i$ . With this notation in place, the goal of an  $n$ -player game is to find an *equilibrium*, a joint action  $\mathbf{x}_* = (x_1^*, \dots, x_n^*) \in \mathbb{R}^D$ , formally expressed as:

$$\underset{\mathbf{x}_* = (x_1^*, \dots, x_n^*) \in \mathbb{R}^D}{\text{find}} \quad f_i(x_*^i; x_*^{-i}) \leq f_i(x^i; x_*^{-i}), \quad \forall x^i \in \mathbb{R}^{d_i}, \quad \forall i \in [n], \quad (1)$$

where  $f_i(x^i; x^{-i}) = f_i(x^1, \dots, x^n)$ .

**MpFL.** As mentioned above, in the setting of interest of this work, we focus on an  $n$ -player game in which multiple players communicate via a central server to reach an equilibrium. In this setting, each player of the  $n$ -player game represents a client to the system (see Figure 1). Mathematically, the problem is formulated as solving (1) with

$$f_i(x^1, \dots, x^n) = \mathbb{E}_{\xi^i \sim \mathcal{D}_i} [f_{i, \xi^i}(x^1, \dots, x^n)].$$

Here  $\mathcal{D}_i$  denotes the data distribution of the  $i$ -th player,  $f_{i, \xi^i}$  is the loss of the  $i$ -th player for a data point  $\xi^i$  sampled from  $\mathcal{D}_i$ .

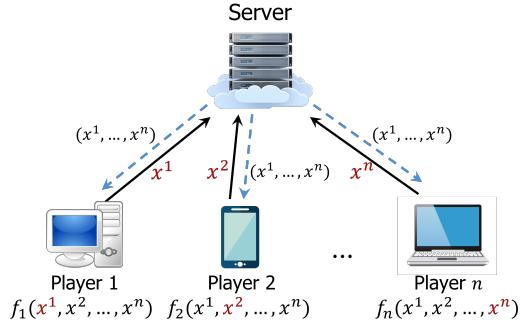


Figure 1: Illustration of MpFL for heterogeneous functions  $f_i$ . The goal is for each player to reach the equilibrium  $\mathbf{x}_* = (x_1^*, \dots, x_n^*)$  (see (1)) with as little communication as possible.

In our proposed FL environment, each client/player uses the strategies of all players to execute local updates. In particular, each player keeps the other players' strategies fixed and updates their own value, which is later shared with the master server, which concatenates all new strategies and sends them back to all players. Later, in Section 3 we introduce and analyze Algorithm 1, named *Per-Player Local SGD* (**PEARL-SGD**), which formalizes the above setting.

Similarly to classical FL, our setting allows for *heterogeneous (non-iid) data* as we have no restrictive assumption on the data distribution  $\mathcal{D}_i$  or the similarity between the functions of the players.

**Assumptions on multiplayer game.** Let us present the main assumptions on the functions of the multiplayer game, which we later use to provide the convergence analysis for **PEARL-SGD**—objective function  $f_i$  of each player  $i \in [n]$  is convex and smooth.

Throughout this work, we denote the partial derivative of  $f_i$  with respect to  $x^i$  as  $\nabla_{x^i} f_i(x^1, \dots, x^n) = \nabla f_i(x^i; x^{-i})$ . This convention allows us to remove the cumbersome subscript  $x^i$  from the  $\nabla$  notation, with the understanding that we only differentiate  $f_i$  with respect to  $x^i$  but never with  $x^{-i}$ .

**Assumption 2.1 (Convex (CVX)).** For  $i \in [n]$ , for any  $x^{-i} \in \mathbb{R}^{D-d_i}$ , the local function  $f_i(\cdot; x^{-i}): \mathbb{R}^{d_i} \rightarrow \mathbb{R}$  is convex. That is, for any  $x^i, y^i \in \mathbb{R}^{d_i}$  and  $x^{-i} \in \mathbb{R}^{D-d_i}$ ,

$$f_i(y^i; x^{-i}) \geq f_i(x^i; x^{-i}) + \langle \nabla f_i(x^i; x^{-i}), y^i - x^i \rangle$$

**Assumption 2.2 (Smoothness (SM)).** For  $i \in [n]$ , for any  $x^{-i} \in \mathbb{R}^{D-d_i}$ , the local function  $f_i(\cdot; x^{-i}): \mathbb{R}^{d_i} \rightarrow \mathbb{R}$  is  $L_i$ -smooth. That is, for any  $x^i, y^i \in \mathbb{R}^{d_i}$  and  $x^{-i} \in \mathbb{R}^{D-d_i}$ ,

$$\|\nabla f_i(x^i; x^{-i}) - \nabla f_i(y^i; x^{-i})\| \leq L_i \|x^i - y^i\|.$$

As in (1), in the stochastic regime of MpFL we have  $f_i(x^1, \dots, x^n) = \mathbb{E}_{\xi^i \sim \mathcal{D}_i} [f_{i, \xi^i}(x^1, \dots, x^n)]$ . To obtain convergence guarantees for **PEARL-SGD** in this scenario, we need the following assumption of bounded variance of the gradient oracle, commonly used in stochastic optimization.

**Assumption 2.3 (Bounded Variance (BV)).** Let  $\sigma_i \geq 0$ ,  $\forall i \in [n]$ . For each  $i = 1, \dots, n$ ,

$$\mathbb{E}_{\xi^i \sim \mathcal{D}_i} \left[ \|\nabla f_{i, \xi^i}(x^i; x^{-i}) - \nabla f_i(x^i; x^{-i})\|^2 \right] \leq \sigma_i^2, \quad \forall x^i \in \mathbb{R}^{d_i}, x^{-i} \in \mathbb{R}^{D-d_i}.$$

## 2.2 Comparison with closely related FL frameworks

Having presented the MpFL setting, let us now provide a concise survey of the most closely related setups: classical FL, federated minimax optimization and personalized FL. We compare each of them with MpFL, and explain in detail in Appendix B *why existing FL algorithms are not applicable to MpFL*. An additional list of related work, explaining the literatures on game theory, distributed Nash equilibrium search, learning in games and the usage of game theory for modeling clients' social behavior in FL, and how MpFL is distinguished from them, is provided in Appendix A.

**Federated learning (FL).** The basic formulation for classical FL is [59]:  $\underset{x \in \mathbb{R}^d}{\text{minimize}} f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ . Here,  $x \in \mathbb{R}^d$  represents the global model parameter,  $f_i(x) = \mathbb{E}_{\xi^i \sim \mathcal{D}_i} [F_i(x, \xi^i)]$  denotes the local objective function at client  $i$ , and  $\mathcal{D}_i$  denotes the data distribution of client  $i$ . The local loss functions  $F_i(x, \xi^i)$  are often of the same form across all clients, but the data distribution  $\mathcal{D}_i$  generally varies, capturing data heterogeneity. The foundational communication-efficient algorithm for this setup is FedAvg (Local SGD), proposed and popularized by [93]. Despite its simplicity, Local SGD has shown empirical success in terms of convergence speed and communication cost. Many works provided theoretical explanation for this performance [127, 29, 128, 63].

Note that in these works on classical FL, clients work in a fully cooperative manner to find  $x_* = \underset{x \in \mathbb{R}^d}{\text{argmin}} f(x)$ , unlike in our proposed MpFL, where the clients serve as players of the game and seek an equilibrium among possibly competing (non-cooperative) objectives.

**Federated minimax optimization.** This is a more recent, federated extension of minimax optimization appearing in many ML applications. There the problem is: minimize  $\max_{y \in \mathbb{R}^{d_y}} \mathcal{L}(x, y) =$

$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(x, y)$ . Here  $n$  is the number of clients and  $\mathcal{L}_i$  represents the local loss function at client  $i$  that depends on both  $x$  and  $y$ . It is defined as  $\mathcal{L}_i(x, y) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [\phi_i(x, y, \xi)]$ , where  $\phi_i(x, y, \xi)$  denotes the loss for the data point  $\xi$ , sampled from the local data distribution  $\mathcal{D}_i$  at client  $i$ . The extension of Local SGD for solving this problem are Local Stochastic Gradient Descent-Ascent (SGDA) [25, 122] or Local Stochastic Extragradient (SEG) [8, 9] algorithms. More recently there was also an approach based on primal-dual updates [20].

While this line of work also studied federated learning in the context of minimax optimization and games, it is totally different from MpFL. The setup assumes that *each FL client adjusts the actions of both players  $x$  and  $y$  of the minimax game*, and does not take the *multiplayer* aspect into account. In contrast, in MpFL, we assume that each client  $i$  is a player of a large-scale multiplayer game who *only adjusts their own action  $x^i$*  in the interest of optimizing their own objective  $f_i$ . In our work, we design the novel **PEARL-SGD** algorithm suitable for MpFL, as the existing Local SGDA and Local SEG methods cannot handle the setup.

**Personalized federated learning.** In personalized FL [34, 129, 50, 49, 26, 130], clients aim to learn models tailored to each local data distribution, while generalizing well on all clients’ data [51]. One formulation of personalized FL is  $\min_{\mathbf{x}=(x^1, \dots, x^n) \in \mathbb{R}^{nd}} \frac{1}{n} \sum_{i=1}^n h_i(x^i) + \phi(x^1, \dots, x^n)$ ; in [50, 51], for

example,  $\phi$  is taken as the model consensus regularizer  $\phi(x^1, \dots, x^n) = \frac{\lambda}{2n} \sum_{i=1}^n \|x^i - \bar{x}\|^2$ . Given that each  $h_i$  and  $\phi$  are convex, its first-order optimality condition is equivalent to the equilibrium condition for the  $n$ -player game where players’ objectives are  $f_i(x^i; x^{-i}) = h_i(x^i) + n\phi(x^1, \dots, x^n)$ . Such formulation of personalized FL, therefore, is an instance of MpFL. On the other hand, there exist personalized FL approaches that are not modeled as a direct subclass of MpFL—e.g., [34, 129] proposed to maintain a separate global model  $w$  along with local models  $x^i$ . Still, it is clear that MpFL is closely related to personalized FL. Importantly, however, the purpose of MpFL is not limited to having personalized models suitable for local data distribution, and encompasses settings where  $x^1, \dots, x^n$  differ in dimensionality or structure.

### 3 PEARL-SGD: Algorithm and Convergence Guarantees

In this section, we introduce and analyze Algorithm 1, named *Per-Player Local SGD* (**PEARL-SGD**), which is suitable for the MpFL setting we described in Section 2.

#### 3.1 Algorithm and assumptions

In **PEARL-SGD**, clients/players of the game run SGD independently in parallel to update their strategy (keeping the strategies  $x^{-i}$  of the other players fixed), which are communicated to other players only once in a while. In more detail, in every round of **PEARL-SGD**, each player  $i \in [n]$  runs  $\tau$  iterations of stochastic gradient descent (SGD) with respect to  $f_i(\cdot, x^{-i})$ , having  $x^{-i}$  fixed to be the information of the other players’ actions obtained from the previous synchronization step. Once each player completes  $\tau$  iterations of SGD (local updates), a synchronization occurs; the central server collects the actions of all players, and then the concatenation of all updated strategies/actions is distributed to all clients/players.

---

#### Algorithm 1 Per-Player Local SGD (**PEARL-SGD**)

---

**Input:** Step-sizes  $\gamma_k > 0$ , Synchronization interval  $\tau \geq 1$ , Number of rounds  $R \geq 1$

**for**  $p = 0, \dots, R - 1$  **do**

- Master collects  $x_{\tau p}^i$  from players  $i \in [n]$
- Master distributes  $\mathbf{x}_{\tau p}$  back to players
- for each players**  $i = 1, \dots, n$  **in parallel do**

  - for**  $k = \tau p, \dots, \tau(p+1) - 1$  **do**

    - Draw  $\xi_k^i \sim \mathcal{D}_i$
    - $g_k^i \leftarrow \nabla f_{i, \xi_k^i}(x_k^i; x_{\tau p}^{-i})$
    - $x_{k+1}^i \leftarrow x_k^i - \gamma_k g_k^i$

  - end for**

- end for**

**end for**

---

**Output:**  $\mathbf{x}_{\tau R} \in \mathbb{R}^D$

---

We emphasize that **PEARL-SGD** and its convergence guarantees hold without any assumption on players’ data distributions  $\mathcal{D}_i$ . That is, functions  $f_i$  can be very different between players, and *the setting is fully heterogeneous*.

Let us note that the synchronization step in **PEARL-SGD** involves transferring a  $D = (d_1 + \dots + d_n)$ -dimensional vector  $\mathbf{x}_{\tau p}$  from the master server to the players. This is an important difference compared to the classical FL (minimization problem), where the dimension of the communication vectors is the same from client to master and from master to client, and it does not scale with  $n$ . While **PEARL-SGD** aims to reduce this overhead compared to its distributed variant ( $\tau = 1$ ) by communicating less frequently (with  $\tau > 1$ ), the high complexity of the synchronization step makes MpFL (and distributed  $n$ -player games in general) more suitable for cross-silo FL setups with relatively small number of organizations and more reliable communication bandwidths. We expect that the potentially expensive communication of high-dimensional vectors  $\mathbf{x}_{\tau p}$  could be addressed by incorporating additional techniques such as gradient compression [4, 10]. This is an orthogonal approach to our proposed local methods, and we leave it for future work.

**Assumptions on the joint gradient operator.** We require some definitions and additional assumptions in order to carry out the theory. Define the joint gradient operator  $\mathbb{F}: \mathbb{R}^D \rightarrow \mathbb{R}^D$  as

$$\mathbb{F}(\mathbf{x}) = (\nabla f_1(x^1; x^{-1}), \dots, \nabla f_n(x^n; x^{-n})).$$

**Assumption 3.1 (Quasi-strong monotonicity (QSM)).** There exists  $\mathbf{x}_* = (x_*^1, \dots, x_*^n) \in \mathbb{R}^D$ , an equilibrium where  $\mathbb{F}(\mathbf{x}_*) = 0$  and  $\mu > 0$  such that  $\forall \mathbf{x} \in \mathbb{R}^D$ ,  $\langle \mathbb{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle \geq \mu \|\mathbf{x} - \mathbf{x}_*\|^2$ .

(QSM) is a concept extending quasi-strong convexity [46, 45] to the context of variational inequality problems (VIPs). This condition has been referred to as different names in the literature, such as strong coherent VIPs [126], VIPs with strong stability condition [95], or the strong Minty variational inequality [28]. It generalizes strong monotonicity, capturing some non-monotone problems. In [89], it was proposed and utilized as an assumption ensuring the convergence of SGDA dynamics in minimax games without the well-known issues of cycling or diverging [97, 22]. Later, it was also used in the analysis of stochastic extragradient method [42] and its single-call variants (optimistic and past stochastic extragradient) [19].

**Assumption 3.2 (Star-cocoercivity (SCO)).**  $\mathbb{F}$  is  $\frac{1}{\ell}$ -star-cocoercive, i.e., there is  $\ell > 0$  such that for any  $\mathbf{x} \in \mathbb{R}^D$ ,  $\langle \mathbb{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle \geq \frac{1}{\ell} \|\mathbb{F}(\mathbf{x})\|^2$ .

(SCO) generalizes the class of coercive operators and, interestingly, can hold for non-Lipschitz operators [89]. This has also been taken as minimal assumption for SGDA analysis in prior work [10]. Note that (QSM) and (SCO) together imply  $\mu \|\mathbf{x} - \mathbf{x}_*\| \leq \|\mathbb{F}(\mathbf{x})\| \leq \ell \|\mathbf{x} - \mathbf{x}_*\|$  for any  $\mathbf{x} \in \mathbb{R}^D$ , which implies  $\mu \leq \ell$ . We call  $\kappa = \ell/\mu \geq 1$  the *condition number* of the problem. In Appendix F, we provide a detailed discussion on the set of our theoretical assumptions and explain connections to other commonly assumed properties in the literature such as cocoercivity, Lipschitzness and monotonicity.

### 3.2 Convergence of **PEARL-SGD**: Deterministic setup

First, we provide the convergence result for **PEARL-SGD** with constant step-size  $\gamma_k \equiv \gamma$  in the full-batch (deterministic) scenario, where there is no noise in the gradient computation. While this is directly recovered as a special case of Theorem 3.4 with  $\sigma_i = 0$ , we state it separately as the deterministic case provides several points of discussion that are worth emphasizing on their own.

**Theorem 3.3.** Assume (CVX), (SM), (QSM) and (SCO). Let  $L_{\max} = \max\{L_1, \dots, L_n\}$ ,  $\kappa = \ell/\mu$  and  $0 < \gamma_k \equiv \gamma \leq \frac{1}{\ell\tau+2(\tau-1)L_{\max}\sqrt{\kappa}}$ . Then the Deterministic **PEARL-SGD** (Algorithm 1 with full-batch) converges with the rate

$$\|\mathbf{x}_{\tau R} - \mathbf{x}_*\|^2 \leq (1 - \gamma\tau\mu\zeta)^R \|\mathbf{x}_0 - \mathbf{x}_*\|^2$$

where  $\zeta = 2 - \gamma\ell\tau - 2(\tau - 1)\gamma L_{\max}\sqrt{\kappa/3} > 0$  (by the choice of  $\gamma$ ).

Theorem 3.3 shows that deterministic **PEARL-SGD** converges linearly to an equilibrium. This is in contrast with the case of local gradient descent in FL setup with heterogeneous data, where convergence is only to a neighborhood of the optimum even in the absence of noise [63], unless further correction mechanism is used [100]. This is because the classical FL problem is modeled as finite sum minimization, whereas our MpFL is modeled as a game, for which the existence of a variationally stable equilibrium is a standard assumption for convergence analysis [14, 95, 89, 96]. In particular, when  $\tau = 1$ , the step-size constraint and the convergence rate of Theorem 3.3 reduces to the analysis of gradient descent-ascent (GDA) under the **(QSM)** and **(SCO)** assumptions from [89], showing that our analysis is tight and consistent with the existing literature.

**Player drift and step-size constraint.** If  $\gamma$  does not appropriately scale down with  $\tau$ , then at each round, players' actions (SGD iterates) converge to minimizers of local functions. We call this phenomenon *player drift*, analogous to client drift in classical FL [61], enforcing the  $\mathcal{O}(1/\tau)$  step-size. In our setting, note that the local minimizers  $x_\star^i(x_{\tau p}^{-i}) := \operatorname{argmin}_{x^i \in \mathbb{R}^{d_i}} f_i(x^i; x_{\tau p}^{-i})$  depend on other players' strategies  $x_{\tau p}^{-i}$ . Due to this dependence, under extreme player drift, **PEARL-SGD** may display undesirable dynamics such as diverging to infinity (which can be checked with simple examples such as the two-player quadratic minimax game  $\min_{u \in \mathbb{R}} \max_{v \in \mathbb{R}} \frac{\mu}{2}u^2 + uv - \frac{\mu}{2}v^2$  with  $\mu < 1$ ). As these features are not typically observed in client drift in classical FL, player drift represents a distinct phenomenon despite some conceptual similarities. Therefore, we consider understanding and mitigating player drift an intriguing direction for future work in MpFL, which may necessitate novel insights that differ from existing approaches to client drift [61, 100].

With the step-size constrained to  $\gamma = \mathcal{O}(1/\tau)$ , communication reduction by **PEARL-SGD** is not observed in the deterministic setting (Theorem 3.3) but is achieved in the stochastic setting—see Section 3.3. A concurrent work [148] proposes the *Decoupled SGD* algorithm, which coincides with our **PEARL-SGD**, and shows that it can be communication-efficient even in the deterministic setup under an additional assumption of *weakly coupled* games (with slightly different main assumptions). We provide a more complete discussion of this in Appendix A.

### 3.3 Convergence of **PEARL-SGD**: Stochastic setup

We now discuss the convergence of **PEARL-SGD** with stochastic gradients. We first present the convergence of **PEARL-SGD** to a neighborhood of the equilibrium  $\mathbf{x}_\star$  given constant step-sizes  $\gamma_k \equiv \gamma$ , and then discuss the communication complexity gain we achieve. Then we present the convergence result using a decreasing step-size selection, showing exact sublinear convergence to  $\mathbf{x}_\star$  (rather than its neighborhood). We provide the outline and details of the proofs in Appendix C.

**Theorem 3.4.** Assume **(CVX)**, **(SM)**, **(BV)**, **(QSM)** and **(SCO)** hold. Let  $0 < \gamma_k \equiv \gamma \leq \frac{1}{\ell\tau+2(\tau-1)L_{\max}\sqrt{\kappa}}$  and denote  $q = L_{\max}/\sqrt{\ell\mu}$ . Then **PEARL-SGD** exhibits the rate:

$$\mathbb{E} \left[ \|\mathbf{x}_{\tau R} - \mathbf{x}_\star\|^2 \right] \leq (1 - \gamma\tau\mu\zeta)^R \|\mathbf{x}_0 - \mathbf{x}_\star\|^2 + \left( 1 + (\tau - 1) \left( (4 + \sqrt{3}q)\gamma\tau L_{\max} + \frac{q}{2\tau} \right) \right) \frac{\gamma\sigma^2}{\mu\zeta}.$$

where  $\sigma^2 = \sum_{i=1}^n \sigma_i^2$  and  $\zeta = 2 - \gamma\ell\tau - 2(\tau - 1)\gamma L_{\max}\sqrt{\kappa/3} > 0$  by the choice of  $\gamma$ .

When  $\tau = 1$ , with  $\gamma \leq 1/\ell$ , the above rate becomes  $\mathbb{E} \left[ \|\mathbf{x}_R - \mathbf{x}_\star\|^2 \right] \leq (1 - \gamma\mu)^R \|\mathbf{x}_0 - \mathbf{x}_\star\|^2 + \gamma\sigma^2/\mu$ , which is consistent with the classical analysis of the stochastic gradient descent-ascent (SGDA). In the result, note that  $\sigma^2$  is the sum of  $\sigma_i^2$ 's, the (upper bounds on) playerwise gradient variances ( $\sigma_i^2 \geq \mathbb{E}_{\xi^i \sim \mathcal{D}_i} \left[ \|\nabla f_i(\xi^i; x^{-i}) - \nabla f_i(x^i; x^{-i})\|^2 \right]$ ). Hence,  $\sigma^2$  represents the upper bound on the variance in estimating the joint gradient operator  $\mathbf{F}(\cdot)$ .

**Remark.** If we use the largest possible step-size  $\gamma = \frac{1}{\ell\tau+2(\tau-1)L_{\max}\sqrt{\kappa}}$  allowed in Theorem 3.4, then the right-hand side of the bound does not scale down indefinitely with  $\tau$ . In fact, with this choice of  $\gamma$ , one can expect the communication gain by a factor of approximately  $\frac{L_{\max}}{\ell}$  (when  $L_{\max} \ll \ell$ ). More precisely, suppose  $q \leq 1$  (equivalently  $L_{\max} \leq \sqrt{\ell\mu}$ —refer to Appendix F for the explanation

that this is a common parameter regime). Then we have  $\gamma = \Theta\left(\frac{1}{\ell\tau}\right)$  and

$$\mathbb{E}\left[\|\mathbf{x}_{\tau R} - \mathbf{x}_*\|^2\right] \leq (1 - \gamma\tau\mu)^R \|\mathbf{x}_0 - \mathbf{x}_*\|^2 + \mathcal{O}\left(\frac{1}{\tau} + \frac{L_{\max}}{\ell}\right) \frac{\sigma^2}{\ell\mu}.$$

The first linear convergence term is essentially unaffected by  $\tau$ , as the effect of using smaller  $\gamma = \Theta\left(\frac{1}{\ell\tau}\right)$  is canceled out by the factor  $\tau$  within  $(1 - \gamma\tau\mu)^R$ . In the second term (which is usually dominant), we see that the size of the convergence neighborhood is reduced by the factor  $\mathcal{O}\left(\frac{1}{\tau} + \frac{L_{\max}}{\ell}\right) = \mathcal{O}\left(\frac{1}{\tau} + \frac{1}{\sqrt{\kappa}}\right)$ . Therefore, we see that with  $\tau = \Omega(\sqrt{\kappa})$ , **PEARL-SGD** reaches about  $\sqrt{\kappa}$  times smaller neighborhood within the same number of communication rounds  $R$  (compared to the case  $\tau = 1$ ).

In Corollary 3.5, we analyze the convergence and communication gain of **PEARL-SGD** in the regime where the total number of iterations  $T = \tau R$  is large, using a step-size depending on  $T$ .

**Corollary 3.5.** Suppose the assumptions of Theorem 3.4 hold, and let  $\tau \geq 1$  be fixed. Let  $q = L_{\max}/\sqrt{\ell\mu}$ . Then **PEARL-SGD** with  $\gamma_k \equiv \gamma = \frac{1}{\mu\eta(1+2q)}$  exhibits the rate

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{(1+q)^2 \|\mathbf{x}_0 - \mathbf{x}_*\|^2}{T^2} + \frac{(1+q)\sigma^2}{\mu^2 T} + \frac{(1+q)\tau^2 L_{\max} \sigma^2}{\mu^3 T^2}\right)$$

where  $\eta$  is selected so that  $T = 2(1+2q)\eta \log \eta$ , provided that  $T$  is large enough so that  $\eta > \kappa\tau$ . Here the  $\tilde{\mathcal{O}}$ -notation hides polylogarithmic terms in  $T$  and constant factors.

**Reduction of communication complexity.** Note that the  $\tilde{\mathcal{O}}\left(\frac{(1+q)^2 \|\mathbf{x}_0 - \mathbf{x}_*\|^2}{T^2}\right)$  term decays fast in Corollary 3.5 (as  $T$  grows) and the terms proportional to  $\sigma^2$  become dominant. The order of convergence is not slower than the  $\tilde{\mathcal{O}}(1/T)$  rate of the fully communicating case  $\tau = 1$ , provided that  $\tau^2 L_{\max} \sigma^2 / \mu^3 T^2 = \mathcal{O}(\sigma^2 / \mu^2 T) \iff \tau = \mathcal{O}\left(\sqrt{\mu T / L_{\max}}\right)$ . Therefore, as long as we select  $\tau = \mathcal{O}\left(\sqrt{\mu T / L_{\max}}\right)$ , in **PEARL-SGD** the communication cost is reduced by the factor of  $\tau$  (because the total number of communications is  $T/\tau$ ). With the largest possible  $\tau$ , the resulting communication complexity is  $T/\tau = \Theta\left(\sqrt{TL_{\max}/\mu}\right) = \Theta\left(\sqrt{T}\right)$ .

**Convergence to equilibrium via decreasing step-sizes.** We conclude the section with convergence result for **PEARL-SGD** using a decreasing step-size selection. While showing a similar convergence rate in terms of  $T$  as in Corollary 3.5, Theorem 3.6 has the advantage of not requiring to fix  $T$  in advance to determine the step-sizes.

**Theorem 3.6.** Under the assumptions of Theorem 3.4, let  $q = L_{\max}/\sqrt{\ell\mu}$ , and choose the step-sizes  $\gamma_k = \begin{cases} \frac{1}{\ell\tau(1+2q)} & \text{if } p < 2(1+2q)\kappa \\ \frac{1}{\tau\mu} \frac{2p+1}{(p+1)^2} & \text{if } p \geq 2(1+2q)\kappa \end{cases}$  for  $\tau p \leq k \leq \tau(p+1) - 1$ ,  $p = 0, \dots, R-1$ . Then **PEARL-SGD** converges with the rate

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}_*\|^2\right] &\leq \frac{4(1+2q)^2 \kappa^2 \tau^2 \|\mathbf{x}_0 - \mathbf{x}_*\|^2}{eT^2} + \frac{4(1+q)\sigma^2}{\mu^2 T} \\ &\quad + \frac{4(1+2q)^2 \kappa \tau \sigma^2}{\mu^2 T^2} \left(1 + \frac{2\tau}{\sqrt{\kappa}}\right) + \frac{32(1+q)\tau^2 L_{\max} \sigma^2 \log T}{\mu^3 T^2} \end{aligned}$$

where  $T = \tau R$  is the total number of iterations.

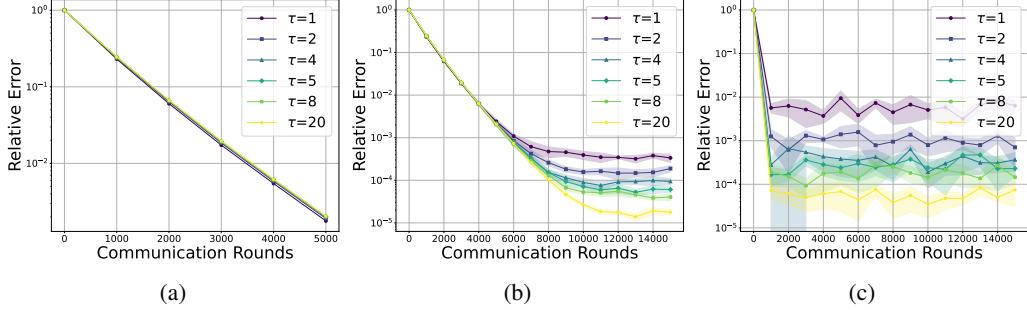


Figure 2: Performance plots for **PEARL-SGD**. Figures 2a (deterministic) and 2b (stochastic) show the relative error  $\frac{\|\mathbf{x}_k - \mathbf{x}_*\|^2}{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}$  on the  $n$ -player game defined by (2) with different values of  $\tau$ , using theoretical step-sizes. (We provide additional experiments for this  $n$ -player game setup in Appendix E.) Figure 2c shows the relative error in the (stochastic) mobile robot control setup (3) for distinct values of  $\tau$ .

## 4 Numerical Experiments

In this section, we conduct experiments to assess the empirical performance of **PEARL-SGD** and verify our theory. We focus on two setups: a multiplayer game with quadratic objectives, and a distributed mobile robot control problem. Details of experiments are provided in Appendix D.

### 4.1 Quadratic $n$ -player game

We consider an  $n$ -player game where the local function of the  $i$ -th player is given by

$$f_i(x^i; x^{-i}) := \frac{1}{M} \sum_{m=1}^M f_{i,m}(x^i; x^{-i}), \quad (2)$$

for  $i = 1, \dots, n$  (with  $d_1 = \dots = d_n = d$ ). In this setting, each  $f_{i,m}(x^i; x^{-i})$  takes the form  $f_{i,m}(x^i; x^{-i}) = \frac{1}{2} \langle x^i, \mathbf{A}_{i,m} x^i \rangle + \sum_{1 \leq j \leq n, j \neq i} \langle x^i, \mathbf{B}_{i,j,m} x^j \rangle + \langle a_{i,m}, x^i \rangle$ , where  $\mathbf{A}_{i,m}, \mathbf{B}_{i,j,m} \in \mathbb{R}^{d \times d}$  and  $a_{i,m} \in \mathbb{R}^d$  for  $m = 1, \dots, M$ .

**Connection to game & control theory literature.** The above  $n$ -player game formulation has been often considered in game and control theory literature and has been used in recent works on distributed games (Nash equilibrium search) [117, 140, 80, 132, 60]. In connection with this literature, in Section 4.2, we demonstrate an experiment on a concrete robot control setup from [60].

We run **PEARL-SGD** with the theoretical step-size  $\gamma = 1/(\ell\tau + 2(\tau-1)L_{\max}\sqrt{\kappa})$  from Theorems 3.3 and 3.4 with  $\tau \in \{1, 2, 4, 5, 8, 20\}$ . We set the cocoercivity parameter to  $\ell = L^2/\mu$  following [33], where  $L$  and  $\mu$  are explicitly computed Lipschitz constant and strong monotonicity parameter of  $\mathbb{F}$ . Figure 2a displays the results from Deterministic **PEARL-SGD**, where we observe that all values of  $\tau$  produce indistinguishable performance plots (which is predicted, as  $\gamma$  scales down with  $\tau$ ). Figure 2b shows results from the stochastic setting (we mini-batch from the finite sum (2)), where we repeat each experiment 5 times and plot the mean relative error with standard deviation (shaded region). It demonstrates that **PEARL-SGD** with larger synchronization interval  $\tau$  provides a clear benefit of achieving smaller relative error  $\frac{\|\mathbf{x}_k - \mathbf{x}_*\|^2}{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}$  using the same number of communication rounds. These results verify our theoretical predictions from Section 3.

In Appendix E.1, we provide additional simulations regarding the case where the precise theoretical parameters  $\mu, \ell, L_{\max}$  are not known, so that  $\gamma$  has to be tuned empirically. It demonstrates that in practice,  $(\tau, \gamma)$  can be effective tunable hyperparameters for gaining communication efficiency.

**Performance of PEARL-SGD for different  $(\gamma, \tau)$  pairs.** Figure 3 displays the heatmap of relative errors (log-scale) after 100 communication rounds of Deterministic **PEARL-SGD** in the case of quadratic game with  $n = 2$ . White and yellow regions indicate divergence/poor performance; darker regions indicate lower relative errors.

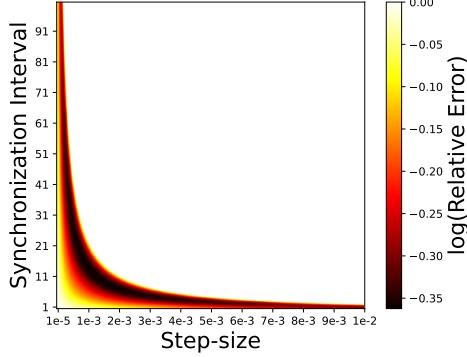


Figure 3: Heatmap of relative errors in logarithmic scale.

Figure 3 reveals a trend—for a fixed  $\gamma$ , **PEARL-SGD**’s performance improves as  $\tau$  increases up to certain threshold, after which it declines and finally diverges. Another key observation is that the dark region of the heatmap (signifying the best performance) takes the shape of a hyperbola. This is consistent with our Theorem 3.3, showing the relationship  $\gamma_\tau \propto 1/\tau$  where  $\gamma_\tau$  is the optimal step-size choice given  $\tau$  (providing fastest convergence).

#### 4.2 Distributed mobile robot control

Here, we consider a distributed control problem of mobile robots from [60]. This is a multi-agent system where each robot has its own objective, depending on the positions  $x^i \in \mathbb{R}^d$  (corresponding to action/strategy in our formulation of multiplayer game) of each  $i$ -th robot. Specifically, the objective function of the robot  $i$  is:

$$f_i(\mathbf{x}) = \underbrace{\frac{a_i}{2} \|x^i - x_{\text{anc}}^i\|^2}_{:= J_{i1}(x^i)} + \underbrace{\frac{b_i}{2} \sum_{j=1}^N \|x^i - x^j - h_{ij}\|^2}_{:= J_{i2}(x^i; x^{-i})} \quad (3)$$

where  $J_{i1}(x^i)$  represents the cost penalizing the distance of agent  $i$  from the anchor point  $x_{\text{anc}}^i \in \mathbb{R}^d$ , and  $J_{i2}(x^i; x^{-i})$  is the cost associated with the relative displacement between the robots’ positions. The control problem finds an equilibrium of the  $n$ -player game, which is the concatenation of all robots’ position vectors, ensuring that each robot stays close to  $x_{\text{anc}}^i$  while maintaining designated displacement from other robots.

We implement **PEARL-SGD** with synchronization intervals  $\tau \in \{1, 2, 4, 5, 8, 20\}$  and the theoretical step-size  $\gamma = \frac{1}{\ell\tau + L_{\max}(\tau-1)\sqrt{\kappa}}$ . Figure 2c shows that with larger values of  $\tau$ , **PEARL-SGD** achieves better accuracy (in terms of distance to  $\mathbf{x}_\star$ ) within a given number of communication rounds. This highlights the potential benefit of using local update steps in solving real-world problems formulated as multiplayer games.

### 5 Conclusion

In this paper, we introduce Multiplayer Federated Learning (MpFL), an FL framework that models setups where clients, acting strategically in their own interests, collaborate through a central server to train models (actions) with the goal of reaching an equilibrium. We propose the **PEARL-SGD** algorithm for handling MpFL and establish its tight convergence guarantees under heterogeneous settings where each player has distinct objectives and data distributions. We show that **PEARL-SGD** achieves improved communication efficiency, mitigating the primary computational bottleneck in large-scale applications.

Our work offers several promising future research directions, including the incorporation of ideas such as Extragradient methods [68, 43], asynchronous updates [23, 127], gradient compression [4], gradient tracking [107], and algorithmic correction for client drifts [61, 100]. We anticipate that our initiation of the study of MpFL will stimulate further research along these and related directions.

## Acknowledgments

Sayantan Choudhury acknowledges support from the Chen Family Dissertation Fellowship. Nicolas Loizou acknowledges support from CISCO Research and the JHU Catalyst Award.

## References

- [1] J. Abernethy, K. A. Lai, and A. Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. *international Conference on Algorithmic Learning Theory*, 2021.
- [2] E. Ahmed and H. Agiza. Dynamics of a cournot game with  $n$ -competitors. *Chaos, Solitons & Fractals*, 9(9):1513–1517, 1998.
- [3] E. Aiyoshi and K. Shimizu. Hierarchical decentralized systems and its new solution by a barrier method. *IEEE Transactions on Systems, Man and Cybernetics*, (6):444–449, 1981.
- [4] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Neural Information Processing Systems*, 2017.
- [5] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23, 2022.
- [6] S. Arisdakessian, O. A. Wahab, A. Mourad, and H. Otrok. Coalitional federated learning: Improving communication and training on non-iid data with selfish clients. *IEEE Transactions on Services Computing*, 16(4):2462–2476, 2023.
- [7] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. *International Conference on Artificial Intelligence and Statistics*, 2020.
- [8] A. Beznosikov, V. Samokhin, and A. Gasnikov. Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*, 2020.
- [9] A. Beznosikov, P. Dvurechenskii, A. Koloskova, V. Samokhin, S. U. Stich, and A. Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Neural Information Processing Systems*, 2022.
- [10] A. Beznosikov, E. Gorbunov, H. Berard, and N. Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. *International Conference on Artificial Intelligence and Statistics*, 2023.
- [11] A. Blum, N. Haghtalab, R. L. Phillips, and H. Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. *International Conference on Machine Learning*, 2021.
- [12] R. I. Boț, E. R. Csetnek, and D.-K. Nguyen. Fast Optimistic Gradient Descent Ascent (OGDA) Method in Continuous and Discrete Time. *Foundations of Computational Mathematics*, 2023.
- [13] J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations research*, 21(1):37–44, 1973.
- [14] M. Bravo, D. Leslie, and P. Mertikopoulos. Bandit learning in concave n-person games. *Neural Information Processing Systems*, 2018.
- [15] G. W. Brown. *Some Notes on Computation of Games Solutions*. Rand Corporation, 1949.
- [16] Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. *International conference on Learning Representations*, 2023.
- [17] A. Capitaine, E. Boursier, A. Scheid, E. Moulines, M. Jordan, E.-M. El-Mhamdi, and A. Durmus. Unravelling in collaborative learning. *Neural Information Processing Systems*, 37: 97231–97260, 2024.

- [18] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. *Neural Information Processing Systems*, 2019.
- [19] S. Choudhury, E. Gorbunov, and N. Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. *Neural Information Processing Systems*, 2024.
- [20] L. Condat and P. Richtárik. RandProx: Primal-dual optimization algorithms with randomized proximal updates. *NeurIPS OPT 2022 Workshop*, 2022.
- [21] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Neural Information Processing Systems*, 2018.
- [22] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. *International Conference on Learning Representations*, 2018.
- [23] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, et al. Large scale distributed deep networks. *Neural Information Processing Systems*, 2012.
- [24] V. Dem'yanov and A. Pevnyi. Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52, 1972.
- [25] Y. Deng and M. Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. *International Conference on Artificial Intelligence and Statistics*, 2021.
- [26] Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive personalized federated learning. *arXiv:2003.13461*, 2020.
- [27] J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. *Conference on Learning Theory*, 2020.
- [28] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. *International Conference on Artificial Intelligence and Statistics*, 2021.
- [29] A. Dieuleveut and K. K. Patel. Communication trade-offs for Local-SGD with large step size. *Neural Information Processing Systems*, 2019.
- [30] K. Donahue and J. Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. *AAAI Conference on Artificial Intelligence*, 2021.
- [31] F. E. Dorner, N. Konstantinov, G. Pashaliev, and M. Vechev. Incentivizing honesty among competitors in collaborative learning and optimization. *Neural Information Processing Systems*, 2023.
- [32] K. Emmanouilidis, R. Vidal, and N. Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. *International Conference on Artificial Intelligence and Statistics*, 2024.
- [33] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.
- [34] A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Neural Information Processing Systems*, 2020.
- [35] S. Feng, B. Li, H. Yu, Y. Liu, and Q. Yang. Semi-supervised federated heterogeneous transfer learning. *Knowledge-Based Systems*, 252:109384, 2022.
- [36] T. Fiez and L. J. Ratliff. Local convergence analysis of gradient descent ascent with finite timescale separation. *International Conference on Learning Representations*, 2021.

- [37] I. Gemp, R. Patel, Y. Bachrach, M. Lanctot, V. Dasagi, L. Marrs, G. Piliouras, S. Liu, and K. Tuyls. Steering language models with game-theoretic solvers. *ICML Agentic Markets Workshop*, 2024.
- [38] A. Giannou, E.-V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. *Neural Information Processing Systems*, 2021.
- [39] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *International conference on Learning Representations*, 2019.
- [40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Neural Information Processing Systems*, 2014.
- [41] E. Gorbunov, F. Hanzely, and P. Richtarik. Local SGD: Unified theory and new efficient methods. *International Conference on Artificial Intelligence and Statistics*, 2021.
- [42] E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. *International Conference on Artificial Intelligence and Statistics*, 2022.
- [43] E. Gorbunov, N. Loizou, and G. Gidel. Extragradient method:  $O(1/k)$  last-iterate convergence for monotone variational inequalities and connections with cocoercivity. *International Conference on Artificial Intelligence and Statistics*, 2022.
- [44] E. Gorbunov, A. Taylor, and G. Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Neural Information Processing Systems*, 2022.
- [45] R. Gower, O. Sebbouh, and N. Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. *International Conference on Artificial Intelligence and Statistics*, 2021.
- [46] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. Sgd: General analysis and improved rates. *International Conference on Machine Learning*, 2019.
- [47] M. Grudzień, G. Malinovsky, and P. Richtarik. Can 5th generation local training methods support client sampling? Yes! *International Conference on Artificial Intelligence and Statistics*, 2023.
- [48] F. Haddadpour and M. Mahdavi. On the convergence of local descent methods in federated learning. *arXiv:1910.14425*, 2019.
- [49] F. Hanzely and P. Richtárik. Federated learning of a mixture of global and local models. *arXiv:2002.05516*, 2020.
- [50] F. Hanzely, S. Hanzely, S. Horváth, and P. Richtarik. Lower bounds and optimal algorithms for personalized federated learning. *Neural Information Processing Systems*, 2020.
- [51] F. Hanzely, B. Zhao, and m. kolar. Personalized federated learning: A unified framework and universal optimization techniques. *Transactions on Machine Learning Research*, 2023.
- [52] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
- [53] J. C. Harsanyi and R. Selten. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.
- [54] Y.-P. Hsieh, P. Mertikopoulos, and V. Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. *International Conference on Machine Learning*, 2021.
- [55] Z. Hu and H. Huang. Tighter analysis for ProxSkip. *International Conference on Machine Learning*, 2023.

[56] M. Huang, D. Zhang, and K. Ji. Achieving linear speedup in non-IID federated bilevel learning. *International Conference on Machine Learning*, 2023.

[57] A. P. Jacob, Y. Shen, G. Farina, and J. Andreas. The consensus game: Language model generation via equilibrium search. *International Conference on Learning Representations*, 2024.

[58] J. Jordan. Three problems in learning mixed-strategy nash equilibria. *Games and Economic Behavior*, 5(3):368–386, 1993.

[59] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[60] D. Kalyva and H. E. Psillakis. Distributed control of a mobile robot multi-agent system for Nash equilibrium seeking with sampled neighbor information. *Automatica*, 166:111712, 2024.

[61] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. *International Conference on Machine Learning*, 2020.

[62] S. P. Karimireddy, W. Guo, and M. I. Jordan. Mechanisms that incentivize data sharing in federated learning. *arXiv:2207.04557*, 2022.

[63] A. Khaled, K. Mishchenko, and P. Richtarik. Tighter theory for local SGD on identical and heterogeneous data. *International Conference on Artificial Intelligence and Statistics*, 2020.

[64] M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar. Adaptive gradient-based meta-learning methods. *Neural Information Processing Systems*, 2019.

[65] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized SGD with changing topology and local updates. *International Conference on Machine Learning*, 2020.

[66] J. Konečný, H. Brendan McMahan, F. X. Yu, P. Richtárik, A. Theertha Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv e-prints*, pages arXiv–1610, 2016.

[67] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016.

[68] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976.

[69] D. M. Kreps. *Game Theory and Economic Modelling*. Oxford University Press, 1990.

[70] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in the finitely repeated prisoners’ dilemma. *Journal of Economic theory*, 27(2):245–252, 1982.

[71] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Neural Information Processing Systems*, 2017.

[72] J. Lee, H. Cho, and C. Yun. Fundamental benefit of alternating updates in minimax optimization. *International Conference on Machine Learning*, 2024.

[73] S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Neural Information Processing Systems*, 2021.

[74] C. J. Li, Y. Yu, N. Loizou, G. Gidel, Y. Ma, N. Le Roux, and M. Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. *International Conference on Artificial Intelligence and Statistics*, 2022.

[75] J. Li, F. Huang, and H. Huang. Communication-efficient federated bilevel optimization with global and local lower level problems. *Neural Information Processing Systems*, 2023.

[76] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2021.

[77] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[78] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Annual Conference on Machine Learning and Systems*, 2020.

[79] T. Li, K. Zhu, N. C. Luong, D. Niyato, Q. Wu, Y. Zhang, and B. Chen. Applications of multi-agent reinforcement learning in future Internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 24(2):1240–1279, 2022.

[80] Z. Li and Z. Ding. Distributed nash equilibrium searching via fixed-time consensus-based algorithms. In *2019 American Control Conference (ACC)*, pages 2765–2770, 2019.

[81] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *International Conference on Machine Learning*, 2020.

[82] T. Lin, Z. Zhou, P. Mertikopoulos, and M. Jordan. Finite-time last-iterate convergence for multi-agent learning in games. *International Conference on Machine Learning*, 2020.

[83] M. Liu, S. Ho, M. Wang, L. Gao, Y. Jin, and H. Zhang. Federated learning meets natural language processing: A survey. *arXiv:2107.12603*, 2021.

[84] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang. FedVision: An online visual object detection platform powered by federated learning. *AAAI Conference on Artificial Intelligence*, 2020.

[85] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4):70–82, 2020.

[86] Y. Liu, X. Zhang, Y. Kang, L. Li, T. Chen, M. Hong, and Q. Yang. FedBCD: A communication-efficient collaborative learning framework for distributed features. *IEEE Transactions on Signal Processing*, 70:4277–4290, 2022.

[87] Y. Liu, Y. Kang, T. Zou, Y. Pu, Y. He, X. Ye, Y. Ouyang, Y.-Q. Zhang, and Q. Yang. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3615–3634, 2024.

[88] N. Loizou. Distributionally robust games with risk-averse players. *International Conference on Operations Research and Enterprise Systems*, 2016.

[89] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Neural Information Processing Systems*, 2021.

[90] R. D. Luce and H. Raiffa. *Games and Decisions: Introduction and Critical Survey*. Courier Corporation, 1989.

[91] O. Marfoq, G. Neglia, A. Bellet, L. Kameni, and R. Vidal. Federated multi-task learning under a mixture of distributions. *Neural Information Processing Systems*, 2021.

[92] E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

[93] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, 2017.

[94] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, 2017.

[95] P. Mertikopoulos and Z. Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019.

[96] P. Mertikopoulos, Y.-P. Hsieh, and V. Cevher. A unified stochastic approximation framework for learning in games. *Mathematical Programming*, 203(1):559–609, 2024.

[97] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of GANs. *Neural Information Processing Systems*, 2017.

[98] J. Mills, J. Hu, and G. Min. Multi-task federated learning for personalised deep neural networks in edge computing. *IEEE Transactions on Parallel and Distributed Systems*, 33(3):630–641, 2021.

[99] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. *International Conference on Artificial Intelligence and Statistics*, 2020.

[100] K. Mishchenko, G. Malinovsky, S. Stich, and P. Richtarik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! *International Conference on Machine Learning*, 2022.

[101] A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Neural Information Processing Systems*, 2021.

[102] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *International Conference on Artificial Intelligence and Statistics*, 2020.

[103] D. Monderer and L. S. Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996.

[104] A. Murhekar, Z. Yuan, B. Ray Chaudhury, B. Li, and R. Mehta. Incentives in federated learning: Equilibria, dynamics, and mechanisms for welfare maximization. *Neural Information Processing Systems*, 2023.

[105] J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

[106] J. F. Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[107] A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[108] Y. Pan and L. Pavel. Global convergence of an iterative gradient algorithm for the nash equilibrium in an extended OSNR game. *IEEE INFOCOM*, 2007.

[109] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[110] Y. Qiu, U. Shanbhag, and F. Yousefian. Zeroth-order methods for nondifferentiable, nonconvex, and hierarchical federated optimization. *Neural Information Processing Systems*, 2023.

[111] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. *Conference on Learning Theory*, 2013.

[112] S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. *Neural Information Processing Systems*, 2013.

[113] J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951.

[114] E. K. Ryu and W. Yin. *Large-scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, 2022.

[115] E. K. Ryu, R. Hannah, and W. Yin. Scaled relative graphs: Nonexpansive operators via 2d euclidean geometry. *Mathematical Programming*, 194(1):569–619, 2022.

[116] W. Saad, Z. Han, H. V. Poor, and T. Basar. Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications. *IEEE Signal Processing Magazine*, 29(5):86–105, 2012.

[117] F. Salehisadaghiani and L. Pavel. Distributed Nash equilibrium seeking: A gossip-based algorithm. *Automatica*, 72:209–216, 2016.

[118] T. C. Schelling. *The Strategy of Conflict: With a New Preface by the Author*. Harvard university press, 1980.

[119] J. Shamma and G. Arslan. Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Transactions on Automatic Control*, 50(3):312–327, 2005.

[120] L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[121] L. S. Shapley et al. *Some topics in two-person games*. Rand Corporation, 1963.

[122] P. Sharma, R. Panda, G. Joshi, and P. Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. *International Conference on Machine Learning*, 2022.

[123] S. Sharma, C. Xing, Y. Liu, and Y. Kang. Secure and efficient federated transfer learning. *IEEE International Conference on Big Data*, 2019.

[124] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. *Neural Information Processing Systems*, 2017.

[125] S. Sokota, R. D’Orazio, J. Z. Kolter, N. Loizou, M. Lanctot, I. Mitliagkas, N. Brown, and C. Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *International Conference on Learning Representations*, 2023.

[126] C. Song, Z. Zhou, Y. Zhou, Y. Jiang, and Y. Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Neural Information Processing Systems*, 2020.

[127] S. U. Stich. Local SGD converges fast and communicates little. *International Conference on Learning Representations*, 2019.

[128] S. U. Stich and S. P. Karimireddy. The error-feedback framework: SGD with delayed gradients. *Journal of Machine Learning Research*, 21(237):1–36, 2020.

[129] C. T. Dinh, N. Tran, and J. Nguyen. Personalized federated learning with moreau envelopes. *Neural Information Processing Systems*, 2020.

[130] A. Z. Tan, H. Yu, L. Cui, and Q. Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2023.

[131] D. A. Tarzanagh, M. Li, C. Thrampoulidis, and S. Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. *International Conference on Machine Learning*, 2022.

[132] T. Tatarenko, W. Shi, and A. Nedić. Geometric convergence of gradient play algorithms for distributed nash equilibrium seeking. *IEEE Transactions on Automatic Control*, 66(11):5342–5353, 2021.

[133] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of Games and Economic Behavior*. Princeton university press, 2007.

[134] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

- [135] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang. Federated learning for healthcare informatics. *Journal of healthcare informatics research*, 5:1–19, 2021.
- [136] J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Neural Information Processing Systems*, 2020.
- [137] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 2019.
- [138] Y. Yang, P. Xiao, and K. Ji. SimFBO: Towards simple, flexible and communication-efficient federated bilevel learning. *Neural Information Processing Systems*, 2023.
- [139] Y. Yang, P. Xiao, S. Ma, and K. Ji. First-order federated bilevel learning. *AAAI Conference on Artificial Intelligence*, 2025.
- [140] M. Ye and G. Hu. Distributed Nash equilibrium seeking by a consensus based approach. *IEEE Transactions on Automatic Control*, 62(9):4811–4818, 2017.
- [141] M. Ye and G. Hu. Game design and analysis for price-based demand response: An aggregate game approach. *IEEE Transactions on Cybernetics*, 47(3):720–730, 2017.
- [142] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with  $O(1/k^2)$  rate on squared gradient norm. *International Conference on Machine Learning*, 2021.
- [143] T. Yoon, J. Kim, J. J. Suh, and E. K. Ryu. Optimal acceleration for minimax and fixed-point problems is not unique. *International Conference on Machine Learning*, 2024.
- [144] H. Yu, S. Yang, and S. Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. *AAAI Conference on Artificial Intelligence*, 2019.
- [145] S. Zhang, S. Choudhury, S. U. Stich, and N. Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. *International Conference on Learning Representations*, 2024.
- [146] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv:1806.00582*, 2018.
- [147] T. Zheng, N. Loizou, P. You, and E. Mallada. Dissipative gradient descent ascent method: A control theory inspired algorithm for min-max optimization. *IEEE Control Systems Letters*, 2024.
- [148] A. Zindari, P. Yazdkhasti, A. Rodomanov, T. Chavdarova, and S. U. Stich. Decoupled sgda for games with intermittent strategy communication. *International Conference on Machine Learning*, 2025.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We explain the novel MpFL framework (Section 2), propose **PEARL-SGD** (Section 3), provide its convergence guarantees (Sections 3.2, 3.3) and experiments (Section 4).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In lines 192–202 and 234–244 we discuss some limitations of the current framework, identifying future research directions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: See Sections [2](#) and [3](#) for the assumptions and Appendix [C](#) for proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: See Appendix [D](#).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: We will include the source codes with instructions in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: See Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We run 5 independent runs for stochastic experiments. Our work does not require further validation of statistical significance as our numerical simulations are primarily conducted for the purpose of verifying the theory.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: See Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#)

Justification: This is a theoretical work and we do not anticipate any direct societal impacts that would require mentioning.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

## 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

# Supplementary Material

We organize the appendix as follows: Section [A](#) provides an additional survey of related work. Section [B](#) provides a detailed explanation on how MpFL differs from prior FL frameworks. Section [C](#) presents the proofs of theoretical results. Section [D](#) provides the details of the experiments omitted from the main paper. Section [E](#) presents some additional experiments. Section [F](#) provides a detailed explanation and interpretation on the theoretical assumptions made in the paper.

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Main contributions	2
<b>2</b>	<b>Multiplayer Federated Learning: Definition and Related Work</b>	<b>2</b>
2.1	Definition of MpFL	3
2.2	Comparison with closely related FL frameworks	4
<b>3</b>	<b>PEARL-SGD: Algorithm and Convergence Guarantees</b>	<b>5</b>
3.1	Algorithm and assumptions	5
3.2	Convergence of PEARL-SGD: Deterministic setup	6
3.3	Convergence of PEARL-SGD: Stochastic setup	7
<b>4</b>	<b>Numerical Experiments</b>	<b>9</b>
4.1	Quadratic $n$ -player game	9
4.2	Distributed mobile robot control	10
<b>5</b>	<b>Conclusion</b>	<b>10</b>
<b>A</b>	<b>Additional related work</b>	<b>28</b>
<b>B</b>	<b>Detailed distinction of MpFL from prior FL setups</b>	<b>30</b>
<b>C</b>	<b>Omitted proofs for Per-Player Local SGD (PEARL-SGD)</b>	<b>31</b>
C.1	Key ideas and proof outline	31
C.2	Proof of Lemma C.1	32
C.3	General properties and bounds for SGD	33
C.4	Proofs of Lemmas C.3 and C.2	37
C.5	Remaining details in proof of Theorem 3.4	38
C.6	Proof of Corollary 3.5	40
C.7	Proof of Theorem 3.6	40
<b>D</b>	<b>Details of Numerical Experiments</b>	<b>42</b>
D.1	Quadratic $n$ -player game	42
D.2	Distributed mobile robot control	42

<b>E Additional Experiments</b>	<b>43</b>
E.1 Quadratic $n$ -player game with step-size tuning . . . . .	43
E.2 Distributed mobile robot control . . . . .	43
<b>F Discussion on Theoretical Assumptions</b>	<b>44</b>
F.1 Possible simplification of assumptions: Assuming cocoercivity of $\mathbb{F}$ . . . . .	44
F.2 Example of non-cocoercive $\mathbb{F}$ satisfying <b>(CVX)</b> , <b>(SM)</b> , <b>(QSM)</b> and <b>(SCO)</b> . . . . .	45

## A Additional related work

**Game Theory & Equilibrium Computation.** Multiplayer games, where multiple players each minimize their own cost function that is affected by the actions of the others, are a long-studied fundamental topic in mathematics and economics [106, 105, 120, 118, 70, 53, 90, 69, 88, 133]. More recently, there has been an increasing interest in the ML community in game-theoretic problems with motivating applications, including adversarial learning [40, 22], multi-agent reinforcement learning (MARL) [71, 79, 125], and language models [37, 57]. This incoming stream of applications has led to the development of novel analyses and insights regarding classical equilibrium-searching algorithms including gradient descent-ascent [24, 81, 136, 36, 147, 89, 72], extragradient [68, 18, 32, 74, 102, 99, 42, 43], optimistic gradient [109, 111, 112, 21, 39, 44] and consensus optimization/Hamiltonian gradient method [97, 7, 1, 89], and even the discovery of new accelerated algorithms for games [27, 142, 73, 16, 12, 143].

**Learning in games.** Without local updates (the case  $\tau = 1$ ), **PEARL-SGD** corresponds to the stochastic gradient play dynamics or the online gradient descent considered in the literature on learning in games [92, 82, 54], or more broadly, regularized Robbins-Monro processes or Follow-the-Generalized-Leader algorithms [96, 38]. Our setup considers the pure Nash equilibrium search in games with continuous and unconstrained action spaces, similarly as in [82]. The theoretical assumptions used in our analysis are similar to the ones that has appeared in this line of works; e.g., [82] uses strong monotonicity and cocoercivity of the joint gradient operator, while we use the weaker notion of quasi-strong monotonicity, which is similar to (but stronger than) the variational stability assumed in multiple works including [14, 95, 96], to name a few. Despite the close connections, our paper is distinguished from these works as we focus on communication efficiency in distributed optimization (federated learning) setup.

**Game theory for social client behavior in FL.** Some prior work have also considered games with strategic clients participating in FL, focusing on designing mechanisms to prevent clients’ social behaviors such as free-riding [104], coalitions [30] or dishonesty [31]. In these works, however, the meaning of the action  $x_i$  is completely different from our setup of interest: it represents the size of the dataset which each client contributes to FL [11, 62, 104], strategy to deceive others and defend against those attacks [31], estimate of the trustworthiness of other clients [6] or willingness to participate in FL [17], all having restrictive meaning in a specific social context. On the other hand, in MpFL,  $x_i$  are clients’ *local models to be optimized through learning*, where the objectives  $f_i$  follow *general game-theoretic structure*, and our primary focus is to design communication-efficient algorithms for finding an equilibrium.

**Heterogeneity and client drift.** One fundamental challenge for theory of Local SGD (FedAvg) is heterogeneity, i.e., varying  $f_i$ ’s due to differences in local data distributions [67, 78]. Under such setup, Local SGD is prone to client drift [146, 61] where local descent trajectories head toward distinct minima (of local objectives), and convergence theories require either additional assumptions [134, 144, 48, 78] or technical analyses [63, 65] to control this drift. Some papers, based on theoretical insights, introduced or analyzed correction mechanisms for Local SGD to mitigate client drift [61, 41, 101, 100, 55, 47]. Extension of such ideas to federated minimax optimization was explored in [145]. We note that the  $n$ -player game setup of MpFL is also fully heterogeneous as each player has distinct (possibly even conflicting) objective functions, and consequently, we have the analogous concept of *player drift*. We refer the readers interested in this topic to the discussion at the end of Section 3.2.

**FL frameworks with individual models.** There are several distinct contexts for FL frameworks (other than personalized FL) where each client learns an individual model. In Vertical FL [137, 86, 87] scenarios, multiple organizations hold distinct features from the common set of samples and they collaborate to train their each local model. In Federated Transfer Learning [123, 85, 35], the participating organizations similarly keep and train local models, but their datasets have heterogeneity over both sample and feature spaces with limited overlaps. Federated Multi-Task Learning [124, 91, 98] extends FL to cases where each client solves different, but related tasks.

**Fictitious play.** The Fictitious Play (FP) is a classical algorithm, originally proposed by [15] to solve minimax games where each player has a finite action space and plays mixed (randomized)

strategies. In FP, each player selects an action that minimizes their expected loss (best response), assuming that the other player plays the empirical (historical) strategy, which is a uniform random mixture of their previously played actions. The convergence of FP to a Nash equilibrium for minimax games was established in [113], but FP fails to converge for general  $n$ -player (with  $n > 2$ ) or non-zero-sum games [121, 58], except for particular cases such as all players having identical objectives [103].

While it may appear that **PEARL-SGD** is conceptually similar to FP (as each player performing multiple local SGD steps can be interpreted as seeking a local approximate best response to others' strategies) the connection is opaque due to some fundamental differences. First, in **PEARL-SGD**, players make their updates based on only the most recent strategies of other players (not the entire history as in FP). Second, in **PEARL-SGD**, local SGD steps are not run until players converge to local optima—this results in player drift as we discuss at the end of Section 3.2, and is rather avoided by using step-sizes scaling down with the number of local steps. Third, in the FP setting players are assumed to have finite action spaces and mixed strategies (corresponding to points on a probability simplex), while the MpFL setting deals with continuous action spaces with pure (non-random) strategies. However, despite distinctions, as FP has been previously studied in the distributed  $n$ -player game setup [119], exploring the further connection between MpFL and FP could be an interesting direction.

**Federated bilevel optimization.** Bilevel optimization is a nested problem in which the outer optimization objective depends on the solution to an inner optimization problem [13, 3]. It can be viewed as a hierarchical game between a leader and a follower and generalizes minimax optimization. Recently, several works have studied bilevel formulations in federated learning (FL) settings [131, 56, 75, 110], with a growing focus on designing communication-efficient, single-loop algorithms for federated bilevel optimization [138, 139].

**Decoupled SGD.** The concurrent work [148] proposes and analyzes the *Decoupled SGD* algorithm whose update rule coincides with **PEARL-SGD**. While the exposition of this work emphasizes Decoupled SGDA — a version for two-player games, the multiplayer case is considered in their Appendix C. The paper defines  $\mathbb{F}_y(x) = (\nabla f_1(x^1; y^{-1}), \dots, \nabla f_n(x^n; y^{-n}))$  for  $y \in \mathbb{R}^D$ , and assuming that  $\mathbb{F}_y$  is  $\bar{\mu}$ -strongly monotone and  $\|\mathbb{F}_y(x) - \mathbb{F}(x)\| \leq L_c \|x - y\|, \forall x, y \in \mathbb{R}^D$ , shows that Decoupled SGD can provide communication acceleration in the *weakly-coupled* regime where  $\kappa_c = L_c/\bar{\mu}$  is small (even in the deterministic regime). Despite the algorithmic similarity between **PEARL-SGD** and Decoupled SGD, our work differs from [148] not only in terms of how technical results are derived, but also in terms of the emphasis. In particular, unlike [148], we consider **PEARL-SGD** as a component of the broader framework of MpFL, which we view as our most significant conceptual contribution.

## B Detailed distinction of MpFL from prior FL setups

**Classical FL algorithms are incompatible with MpFL.** We first clarify that classical FL algorithms including Local SGD [94], FedProx [78], SCAFFOLD [61] or Scaffnew [100] are not suitable for the MpFL setting. There the problem is formulated as

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x). \quad (\text{FL})$$

On the other hand, recall that MpFL is formulated as

$$\underset{\mathbf{x}_* = (x_1^*, \dots, x_n^*) \in \mathbb{R}^D}{\text{find}} \quad f_i(x_*^i; x_*^{-i}) \leq f_i(x^i; x_*^{-i}), \quad \forall x^i \in \mathbb{R}^{d_i}, \quad \forall i \in [n]. \quad (\text{MpFL})$$

One obvious distinction is that (FL) seeks a single  $x_* \in \mathbb{R}^d$  minimizing the finite sum, while in (MpFL) each player finds distinct  $x^1, \dots, x^n$  satisfying the equilibrium condition. The player  $i$  **does not** update  $x^j$  for  $j \neq i$  in MpFL, so unlike in classical FL, each player only contributes to partial coordinates of the desired solution  $\mathbf{x}_* = (x_1^*, \dots, x_n^*)$ . Therefore, we cannot apply Local SGD or its variants to the MpFL setup.

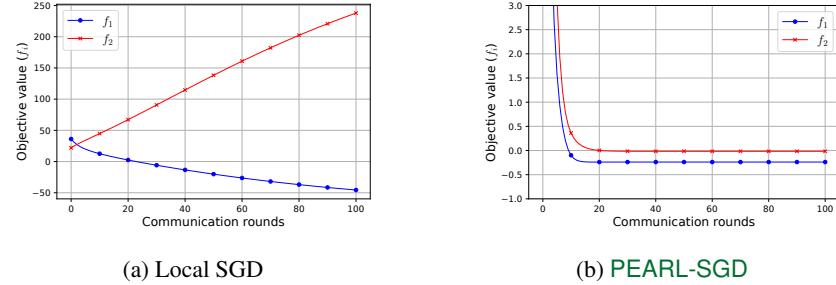


Figure 4: Plots of objective values  $f_1, f_2$  in (4) from running (left) Local SGD on the joint variable  $(u, v)$  and (right) PEARL-SGD.

Additionally, it is generally not possible to reach an equilibrium by performing SGD on the sum of objectives (as in Local SGD). Consider the following simple example with  $n = 2$  clients:

$$f_1(u; v) = \frac{1}{2} u^\top (\mathbf{A}u - a - \mathbf{B}^\top v) - \frac{1}{20} \|v\|^2, \quad f_2(v; u) = \frac{1}{4} \|v\|^2 + \frac{1}{2} v^\top (\mathbf{B}u - b) - \frac{1}{20} \|u\|^2 \quad (4)$$

where  $u, v \in \mathbb{R}^d$ ,  $\mathbf{A} \succ 0$  and  $a, b \in \mathbb{R}^d$  (we use  $(u, v)$  instead of  $(x^1, x^2)$  for clearer notation). Running Local SGD with respect to the joint variable  $(u, v)$  on the sum  $\frac{1}{2}(f_1(u, v) + f_2(u, v))$  results in divergence of one of the objective values (Fig. 4a) while PEARL-SGD converges to the equilibrium and the objective values stabilize (Fig. 4b). However, although we included this example for illustration, note that here Local SGD is not even conforming to the rules of MpFL as both clients are updating  $(u, v)$  simultaneously.

**Federated minimax optimization (FMO) algorithms are incompatible with MpFL.** In FMO, the problem is

$$\underset{x \in \mathbb{R}^{d_x}}{\text{minimize}} \quad \underset{y \in \mathbb{R}^{d_y}}{\text{maximize}} \quad \mathcal{L}(x, y) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(x, y). \quad (\text{FMO})$$

In algorithms designed for (FMO) including Local SGDA or Local SEG, each client  $i$  locally updates *both* the minimization and maximization variables  $(x^i, y^i)$ , and *all clients work collaboratively* toward finding the minimax solution (saddle point) of  $\mathcal{L}(x, y)$ , which is a global objective. On the other hand, in (MpFL), each player  $i$  locally updates *only* their own action  $x^i$ , *for their own individual interest* of reducing  $f_i(\cdot; x^{-i})$ . Even though both (FMO) and (MpFL) aim to reach an equilibrium, they are completely different processes from a conceptual level. For example, for the  $n$ -player game setup in Section 4.1 with  $n = 5$ , it is not possible to apply Local SGDA or Local SEG due to the conceptual mismatch.

## C Omitted proofs for Per-Player Local SGD (PEARL-SGD)

### C.1 Key ideas and proof outline

We first provide an outline for the proof of Theorem 3.4. The key components of the proof are as follows: (i) we show that a round of local SGD in PEARL-SGD behaves like a large single descent step with respect to the joint gradient operator  $\mathbb{F}$  except for *local error* terms caused by running multiple SGD steps locally (Lemma C.1), and (ii) we bound these local error terms (Lemma C.2).

**Lemma C.1.** Assume **(SM)**, and let  $L_{\max} = \max\{L_1, \dots, L_n\}$ . Let  $0 \leq p \leq R-1$  be a fixed round index in PEARL-SGD and suppose  $\gamma_k \equiv \gamma > 0$  for  $k = \tau p, \dots, \tau(p+1)-1$ . Then for arbitrary  $\alpha > 0$ , we have

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \|^2 \mid \mathbf{x}_{\tau p} \right] &\leq (1 + (\tau-1)\alpha\gamma) \|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 - 2\gamma\tau \langle \mathbb{F}(\mathbf{x}_{\tau p}), \mathbf{x}_{\tau p} - \mathbf{x}_* \rangle \\ &\quad + \frac{\gamma L_{\max}^2}{\alpha} \sum_{j=\tau p+1}^{\tau p+\tau-1} \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_j \|^2 \mid \mathbf{x}_{\tau p} \right] + \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_{\tau(p+1)} \|^2 \mid \mathbf{x}_{\tau p} \right]. \end{aligned}$$

**Local error bound.** The right hand side of the bound in Lemma C.1 involves the quantities

$$\mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_j \|^2 \mid \mathbf{x}_{\tau p} \right] = \sum_{i=1}^n \mathbb{E} \left[ \|x_{\tau p}^i - x_j^i\|^2 \mid \mathbf{x}_{\tau p} \right] \quad (5)$$

for  $j = \tau p + 1, \dots, \tau(p+1)$ . We further bound (5) using the following result.

**Lemma C.2.** Suppose Assumptions **(CVX)**, **(SM)** and **(BV)** hold. For a fixed  $i \in [n]$  and a fixed communication round  $p$  in PEARL-SGD, suppose  $\gamma_k \equiv \gamma$  for  $k = \tau p, \dots, \tau(p+1)-1$ , where  $0 < \gamma \leq \frac{1}{L_i} \min \left\{ 1, \frac{1}{\tau-1} \right\}$ . Then for  $t = 0, \dots, \tau$ ,

$$\mathbb{E} \left[ \|x_{\tau p}^i - x_{\tau p+t}^i\|^2 \mid \mathbf{x}_{\tau p} \right] \leq \gamma^2 t^2 \|\nabla f(x_{\tau p}^i; x_{\tau p}^{-i})\|^2 + \gamma^2 t (1 + 2(t-1)(t+1)\gamma L_i) \sigma_i^2.$$

Here we sketch the proof of Lemma C.2 and clarify the role of Assumption **(CVX)**. By assuming that each  $f_i(\cdot; x_{\tau p}^{-i})$  is convex and  $L_i$ -smooth, we can prove Lemma C.3, showing that the expectation of squared gradient norm is “almost” nonincreasing along the local SGD steps, except for some additional term due to stochasticity. Then, we rewrite each summand in (5) as

$$\mathbb{E} \left[ \|x_{\tau p}^i - x_k^i\|^2 \mid \mathbf{x}_{\tau p} \right] = \mathbb{E} \left[ \gamma^2 \left\| \sum_{j=\tau p}^{k-1} g_j^i \right\|^2 \mid \mathbf{x}_{\tau p} \right] = \mathbb{E} \left[ \gamma^2 \left\| \sum_{j=\tau p}^{k-1} \nabla f_{i,\xi_j^i}(x_j^i; x_{\tau p}^{-i}) \right\|^2 \mid \mathbf{x}_{\tau p} \right] \quad (6)$$

and use Lemma C.3 to bound (6).

**Lemma C.3.** Under the assumptions of Lemma C.2, for  $j = \tau p + 1, \dots, \tau(p+1)$ ,

$$\mathbb{E} \left[ \|\nabla f_i(x_j^i; x_{\tau p}^{-i})\|^2 \mid \mathbf{x}_{\tau p} \right] \leq \|\nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i})\|^2 + 2(j-\tau p)\gamma L_i \sigma_i^2.$$

**Remark.** Given (6), it is tempting to apply Jensen’s inequality to the rightmost quantity and then apply Lemma C.3. However, this results in a bound that is looser than our Lemma C.2. We need more careful arguments regarding the expectations, which we detail throughout Appendix C.

*Proof outline for Theorem 3.4.* We combine Lemmas C.1 and C.2, and then apply (SCO) to eliminate the  $\|\mathbf{F}(\mathbf{x}_{\tau p})\|^2$  terms to obtain

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \|^2 \mid \mathbf{x}_{\tau p} \right] &\leq (1 + (\tau - 1)\alpha\gamma) \|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 + (\text{terms proportional to } \sigma^2) \\ &\quad - \underbrace{\left( 2\gamma\tau - \gamma^2\tau^2\ell - \frac{\gamma^3 L_{\max}^2 \tau^2 (\tau - 1)\ell}{3\alpha} \right)}_{:=C} \langle \mathbf{F}(\mathbf{x}_{\tau p}), \mathbf{x}_{\tau p} - \mathbf{x}_* \rangle \quad (7) \end{aligned}$$

Provided that  $C \geq 0$ , we can upper bound the second line of (7) by  $-C\mu \|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2$  using (QSM). Then we choose  $\alpha = \gamma\tau L_{\max} \sqrt{\frac{\ell\mu}{3}}$  which minimizes the resulting coefficient of  $\|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2$ , and rewrite it in the form  $1 - \gamma\tau\mu\zeta$ . Finally, take expectation over  $\mathbf{x}_{\tau p}$  in (7) and unroll the recursion.  $\square$

The proofs of Lemmas C.1, C.2 and C.3 and the detailed full proof of Theorem 3.4 are presented through the following subsections.

## C.2 Proof of Lemma C.1

Note that for  $k = \tau p + 1, \dots, \tau(p+1)$  (iterations between  $p$ -th and  $(p+1)$ -th communications), we have

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}_*\|^2 &= \sum_{i=1}^n \|x_k^i - x_*^i\|^2 \\ &= \sum_{i=1}^n \|x_{\tau p}^i - x_*^i - (x_{\tau p}^i - x_k^i)\|^2 \\ &= \sum_{i=1}^n \left[ \|x_{\tau p}^i - x_*^i\|^2 - 2 \langle x_{\tau p}^i - x_*^i, x_{\tau p}^i - x_k^i \rangle + \|x_{\tau p}^i - x_k^i\|^2 \right] \\ &= \|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 - 2\gamma \sum_{i=1}^n \sum_{j=\tau p}^{k-1} \langle x_{\tau p}^i - x_*^i, g_j^i \rangle + \sum_{i=1}^n \|x_{\tau p}^i - x_k^i\|^2, \quad (8) \end{aligned}$$

where for the last equality, we use

$$g_j^i = \nabla f_{i,\xi_j^i}(x_j^i; x_{\tau p}^{-i}), \quad j = \tau p, \dots, k-1, \quad i = 1, \dots, n$$

and

$$x_{j+1}^i = x_j^i - \gamma g_j^i, \quad j = \tau p, \dots, k-1, \quad i = 1, \dots, n$$

to rewrite  $x_{\tau p}^i - x_k^i = \gamma \sum_{j=\tau p}^{k-1} g_j^i$ . Note that we have

$$\mathbb{E} \left[ -\langle x_{\tau p}^i - x_*^i, g_{\tau p}^i \rangle \mid \mathbf{x}_{\tau p} \right] = -\langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \rangle,$$

while for the other indices  $j = \tau p + 1, \dots, k-1$ , we have the upper bound

$$\begin{aligned} \mathbb{E} \left[ -\langle x_{\tau p}^i - x_*^i, g_j^i \rangle \mid x_j^i \right] &= -\langle x_{\tau p}^i - x_*^i, \nabla f_i(x_j^i; x_{\tau p}^{-i}) \rangle \\ &= -\langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \rangle + \langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) - \nabla f_i(x_j^i; x_{\tau p}^{-i}) \rangle \\ &\leq -\langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \rangle + \frac{\alpha}{2} \|x_{\tau p}^i - x_*^i\|^2 + \frac{1}{2\alpha} \|\nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) - \nabla f_i(x_j^i; x_{\tau p}^{-i})\|^2 \\ &\leq -\langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \rangle + \frac{\alpha}{2} \|x_{\tau p}^i - x_*^i\|^2 + \frac{L_i^2}{2\alpha} \|x_{\tau p}^i - x_j^i\|^2 \end{aligned}$$

where in the fourth line, we use Young's inequality with an arbitrary  $\alpha > 0$  that we determine later. Take expectations of the both sides in (8) (conditioned on  $\mathbf{x}_{\tau p}$ ), and apply the above bound with the

tower rule to obtain

$$\begin{aligned}
& \mathbb{E} \left[ \|\mathbf{x}_k - \mathbf{x}_* \|^2 \mid \mathbf{x}_{\tau p} \right] \\
& \leq \|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 - 2\gamma \sum_{i=1}^n \sum_{j=\tau p}^{k-1} \langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \rangle + 2\gamma \sum_{i=1}^n \sum_{j=\tau p+1}^{k-1} \frac{\alpha}{2} \|x_{\tau p}^i - x_*^i\|^2 \\
& + 2\gamma \sum_{i=1}^n \sum_{j=\tau p+1}^{k-1} \mathbb{E} \left[ \frac{L_i^2}{2\alpha} \|x_{\tau p}^i - x_j^i\|^2 \mid \mathbf{x}_{\tau p} \right] + \sum_{i=1}^n \mathbb{E} \left[ \|x_{\tau p}^i - x_k^i\|^2 \mid \mathbf{x}_{\tau p} \right]. \tag{9}
\end{aligned}$$

Now we apply the identities

$$\begin{aligned}
\sum_{i=1}^n \langle x_{\tau p}^i - x_*^i, \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \rangle &= \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle, \quad \sum_{i=1}^n \|x_{\tau p}^i - x_*^i\|^2 = \|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 \\
\sum_{i=1}^n \mathbb{E} \left[ \|x_{\tau p}^i - x_k^i\|^2 \mid \mathbf{x}_{\tau p} \right] &= \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_k\|^2 \mid \mathbf{x}_{\tau p} \right]
\end{aligned}$$

and the inequality

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=\tau p+1}^{k-1} \mathbb{E} \left[ \frac{L_i^2}{2\alpha} \|x_{\tau p}^i - x_j^i\|^2 \mid \mathbf{x}_{\tau p} \right] &\leq \frac{L_{\max}^2}{2\alpha} \sum_{j=\tau p+1}^{k-1} \sum_{i=1}^n \mathbb{E} \left[ \|x_{\tau p}^i - x_j^i\|^2 \mid \mathbf{x}_{\tau p} \right] \\
&= \frac{L_{\max}^2}{2\alpha} \sum_{j=\tau p+1}^{k-1} \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_j\|^2 \mid \mathbf{x}_{\tau p} \right]
\end{aligned}$$

to (9) and plug in  $k = \tau(p+1)$ , which gives the desired result.

### C.3 General properties and bounds for SGD

In this section, we present some general properties of stochastic gradient descent (SGD) for an  $L$ -smooth, convex function  $f: \mathbb{R}^m \rightarrow \mathbb{R}$ . Suppose that we have a stochastic oracle  $\nabla f_\xi(\cdot)$  for the gradient operator  $\nabla f(\cdot)$ , satisfying

$$\mathbb{E}_\xi[\nabla f_\xi(x)] = \nabla f(x), \quad \mathbb{E}_\xi \left[ \|\nabla f_\xi(x) - \nabla f(x)\|^2 \right] \leq \rho^2, \quad \forall x \in \mathbb{R}^m \tag{10}$$

where  $\mathbb{E}_\xi$  denotes the expectation with respect to randomness in  $\xi$ . This setup and the subsequent results are the abstractions of intermediate results that we need for the proofs of Lemma C.3 and Lemma C.2. Specifically, we will later use the results of this section with

$$f(\cdot) = f_i(\cdot; x_{\tau p}^{-i}), \quad \rho^2 = \sigma_i^2,$$

for each  $i = 1, \dots, n$ . We make this abstraction to simplify notations and to more effectively convey the key intuitions underlying the analyses.

**Lemma C.4.** Let  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  be convex and  $L$ -smooth. Suppose that a stochastic gradient oracle  $\nabla f_\xi(\cdot)$  satisfies (10). Let  $y = x - \gamma \nabla f_\xi(x)$ , where  $0 < \gamma \leq \frac{2}{L}$ . Then we have

$$\mathbb{E}_\xi \left[ \|\nabla f(y)\|^2 \right] \leq \|\nabla f(x)\|^2 + 2\gamma L \rho^2.$$

*Proof.* It is well-known that if  $f$  is convex and  $L$ -smooth, then  $\nabla f$  is  $\frac{1}{L}$ -cocoercive, i.e., for any  $x, y \in \mathbb{R}^m$ ,

$$\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2.$$

By cocoercivity and the step-size condition  $\gamma \leq \frac{2}{L}$ , we have

$$\begin{aligned} & \frac{\gamma}{2} \|\nabla f(x) - \nabla f(y)\|^2 \\ & \leq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \\ & \leq \langle x - y, \nabla f(x) - \nabla f(y) \rangle \\ & = \langle \gamma \nabla f_\xi(x), \nabla f(x) - \nabla f(y) \rangle \\ & = \gamma (\langle \nabla f_\xi(x), \nabla f(x) \rangle - \langle \nabla f(x), \nabla f(y) \rangle + \langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) \rangle). \end{aligned}$$

Taking expectation of the both sides, we obtain

$$\begin{aligned} & \mathbb{E}_\xi \left[ \frac{\gamma}{2} \|\nabla f(x) - \nabla f(y)\|^2 \right] \\ & \leq \mathbb{E}_\xi [\gamma \langle \nabla f_\xi(x), \nabla f(x) \rangle - \gamma \langle \nabla f(x), \nabla f(y) \rangle + \gamma \langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) \rangle] \\ & = \gamma \|\nabla f(x)\|^2 - \gamma \mathbb{E}_\xi [\langle \nabla f(x), \nabla f(y) \rangle] + \gamma \mathbb{E}_\xi [\langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) \rangle]. \end{aligned}$$

Cancelling out the terms and dividing both sides by  $\frac{\gamma}{2}$ , we then have

$$\mathbb{E}_\xi [\|\nabla f(y)\|^2] \leq \|\nabla f(x)\|^2 + 2\mathbb{E}_\xi [\langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) \rangle]. \quad (11)$$

Now observe that

$$\mathbb{E}_\xi [\langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) \rangle] = \mathbb{E}_\xi [\langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) - \nabla f(x - \gamma \nabla f(x)) \rangle]$$

because  $\nabla f(x - \gamma \nabla f(x))$  is a non-random quantity and  $\mathbb{E}_\xi [\nabla f(x) - \nabla f_\xi(x)] = 0$ . Then

$$\begin{aligned} & \mathbb{E}_\xi [\langle \nabla f(x) - \nabla f_\xi(x), \nabla f(y) - \nabla f(x - \gamma \nabla f(x)) \rangle] \\ & = \mathbb{E}_\xi [\langle \nabla f(x) - \nabla f_\xi(x), \nabla f(x - \gamma \nabla f_\xi(x)) - \nabla f(x - \gamma \nabla f(x)) \rangle] \\ & \leq \mathbb{E}_\xi [\|\nabla f(x) - \nabla f_\xi(x)\| \|\nabla f(x - \gamma \nabla f_\xi(x)) - \nabla f(x - \gamma \nabla f(x))\|] \\ & \leq \mathbb{E}_\xi [\|\nabla f(x) - \nabla f_\xi(x)\| L \|(x - \gamma \nabla f_\xi(x)) - (x - \gamma \nabla f(x))\|] \\ & = \gamma L \mathbb{E}_\xi [\|\nabla f(x) - \nabla f_\xi(x)\|^2] \\ & = \gamma L \rho^2, \end{aligned}$$

and plugging this into (11) completes the proof.  $\square$

**Lemma C.5.** Let  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  be convex and  $L$ -smooth and let the stochastic gradient oracle  $\nabla f_\xi(\cdot)$  satisfy (10). Let  $x_0 \in \mathbb{R}^m$  be any initial point,  $0 < \gamma \leq \frac{2}{L}$ , and  $x_1, \dots, x_t$  be a sequence generated by the stochastic gradient descent algorithm

$$x_{s+1} = x_s - \gamma \nabla f_{\xi_s}(x_s)$$

for  $s = 0, \dots, t-1$ . Then we have

$$\mathbb{E} [\|\nabla f(x_s)\|^2] \leq \|\nabla f(x_0)\|^2 + 2s\gamma L\rho^2$$

for  $s = 0, \dots, t-1$ .  $\square$

*Proof.* Apply Lemma C.4 recursively and use the tower rule (law of total expectation).  $\square$

**Lemma C.6.** Let  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  be  $L$ -smooth and let  $x_0, \dots, x_t$  be a sequence generated by stochastic gradient descent

$$x_{s+1} = x_s - \gamma \nabla f_{\xi_s}(x_s)$$

where the stochastic gradient oracle satisfies (10). Let  $\hat{x}_0, \dots, \hat{x}_t$  be generated via *deterministic* gradient descent

$$\hat{x}_{s+1} = \hat{x}_s - \gamma \nabla f(\hat{x}_s)$$

where  $\hat{x}_0 = x_0$ . Then, provided that  $0 < \gamma \leq \frac{1}{L(t-1)}$ , we have

$$\|x_t - \hat{x}_t\| \leq 3\gamma \sum_{s=0}^{t-1} \|\nabla f_{\xi_s}(x_s) - \nabla f(x_s)\|.$$

**Remark.** This result only assumes  $L$ -smoothness of  $f$  (which is  $L$ -Lipschitz continuity of  $\nabla f$ ) and does not require convexity.

*Proof.* When  $t = 1$ , we have  $\|x_t - \hat{x}_t\| = \gamma \|\nabla f_{\xi_0}(x_0) - \nabla f(x_0)\|$  as  $x_0 = \hat{x}_0$ .

Now assume  $t > 1$ . Observe that

$$\begin{aligned} x_t - \hat{x}_t &= (x_{t-1} - \hat{x}_{t-1}) - \gamma (\nabla f_{\xi_{t-1}}(x_{t-1}) - \nabla f(\hat{x}_{t-1})) \\ &= (x_{t-1} - \hat{x}_{t-1}) - \gamma (\nabla f_{\xi_{t-1}}(x_{t-1}) - \nabla f(x_{t-1})) - \gamma (\nabla f(x_{t-1}) - \nabla f(\hat{x}_{t-1})) \end{aligned}$$

and therefore,

$$\begin{aligned} \|x_t - \hat{x}_t\| &\leq \|x_{t-1} - \hat{x}_{t-1}\| + \gamma \|\nabla f_{\xi_{t-1}}(x_{t-1}) - \nabla f(x_{t-1})\| + \gamma \|\nabla f(x_{t-1}) - \nabla f(\hat{x}_{t-1})\| \\ &\leq (1 + \gamma L) \|x_{t-1} - \hat{x}_{t-1}\| + \gamma \|\nabla f_{\xi_{t-1}}(x_{t-1}) - \nabla f(x_{t-1})\| \end{aligned}$$

where the last inequality uses the  $L$ -smoothness assumption. Now unrolling the recursion and using the fact  $\|x_0 - \hat{x}_0\| = 0$  we obtain

$$\begin{aligned} \|x_t - \hat{x}_t\| &\leq \sum_{s=0}^{t-1} \gamma (1 + \gamma L)^{t-s-1} \|\nabla f_{\xi_s}(x_s) - \nabla f(x_s)\| \\ &\leq \gamma \left(1 + \frac{1}{t-1}\right)^{t-1} \sum_{s=0}^{t-1} \|\nabla f_{\xi_s}(x_s) - \nabla f(x_s)\| \\ &\leq 3\gamma \sum_{s=0}^{t-1} \|\nabla f_{\xi_s}(x_s) - \nabla f(x_s)\|. \end{aligned}$$

□

**Lemma C.7.** Under the assumptions of Lemma C.6, we have

$$\mathbb{E} [\langle \nabla f_{\xi_0}(x_0) - \nabla f(x_0), \nabla f(x_t) \rangle] \leq 3t\gamma L\rho^2.$$

*Proof.* Observe that because  $\hat{x}_t$  as defined in Lemma C.6 is a non-random quantity and

$$\mathbb{E} [\nabla f_{\xi_0}(x_0) - \nabla f(x_0)] = 0,$$

we have

$$\mathbb{E} [\langle \nabla f_{\xi_0}(x_0) - \nabla f(x_0), \nabla f(\hat{x}_t) \rangle] = 0.$$

Using this, we can proceed as in the following to obtain the desired bound:

$$\begin{aligned} &\mathbb{E} [\langle \nabla f_{\xi_0}(x_0) - \nabla f(x_0), \nabla f(x_t) \rangle] \\ &= \mathbb{E} [\langle \nabla f_{\xi_0}(x_0) - \nabla f(x_0), \nabla f(x_t) - \nabla f(\hat{x}_t) \rangle] \\ &\leq \mathbb{E} [\|\nabla f_{\xi_0}(x_0) - \nabla f(x_0)\| \|\nabla f(x_t) - \nabla f(\hat{x}_t)\|] \\ &\leq \mathbb{E} [\|\nabla f_{\xi_0}(x_0) - \nabla f(x_0)\| L \|x_t - \hat{x}_t\|] \\ &\leq 3\gamma L \mathbb{E} \left[ \|\nabla f_{\xi_0}(x_0) - \nabla f(x_0)\| \sum_{s=0}^{t-1} \|\nabla f_{\xi_s}(x_s) - \nabla f(x_s)\| \right] \\ &\leq 3\gamma L \mathbb{E} \left[ \sum_{s=0}^{t-1} \left( \frac{\|\nabla f_{\xi_0}(x_0) - \nabla f(x_0)\|^2}{2} + \frac{\|\nabla f_{\xi_s}(x_s) - \nabla f(x_s)\|^2}{2} \right) \right] \\ &\leq 3t\gamma L\rho^2. \end{aligned}$$

□

**Lemma C.8.** Under the assumptions of Lemma C.6, we have

$$\mathbb{E} \left[ \|x_0 - x_t\|^2 \right] \leq \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^{t-1} \nabla f(x_s) \right\|^2 \right] + \gamma^2 t \rho^2 + (t-1)t(t+1)\gamma^3 L \rho^2.$$

*Proof.* In the case  $t = 1$ , we have

$$\mathbb{E} \left[ \|x_0 - x_1\|^2 \right] = \gamma^2 \mathbb{E}_{\xi_0} \left[ \|\nabla f_{\xi_0}(x_0)\|^2 \right] \leq \gamma^2 \rho^2 + \gamma^2 \|\nabla f(x_0)\|^2,$$

which is the desired statement. Now we use induction on  $t$ . Suppose that the result holds for any initial point and  $t$  steps of SGD. Consider a sequence  $x_0, \dots, x_{t+1}$  generated via SGD with initial point  $x_0$  and step-size  $\gamma > 0$ . Observe that

$$\begin{aligned} & \mathbb{E} \left[ \|x_0 - x_{t+1}\|^2 \right] \\ &= \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^t \nabla f_{\xi_s}(x_s) \right\|^2 \right] \\ &= \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^{t-1} \nabla f_{\xi_s}(x_s) \right\|^2 + \mathbb{E}_{\xi_t} \left[ 2 \left\langle \nabla f_{\xi_t}(x_t), \sum_{s=0}^{t-1} \nabla f_{\xi_s}(x_s) \right\rangle + \|\nabla f_{\xi_t}(x_t)\|^2 \mid x_t \right] \right] \\ &\leq \mathbb{E} \left[ \|x_0 - x_t\|^2 \right] + \gamma^2 \mathbb{E} \left[ 2 \left\langle \nabla f(x_t), \sum_{s=0}^{t-1} \nabla f_{\xi_s}(x_s) \right\rangle + \|\nabla f(x_t)\|^2 + \rho^2 \right] \end{aligned} \quad (12)$$

where the third line uses the tower rule. Now observe that for  $s = 0, \dots, t-1$ ,

$$\begin{aligned} \mathbb{E} [\langle \nabla f(x_t), \nabla f_{\xi_s}(x_s) \rangle] &= \mathbb{E} [\langle \nabla f(x_t), \nabla f(x_s) \rangle] + \mathbb{E} [\langle \nabla f(x_t), \nabla f_{\xi_s}(x_s) - \nabla f(x_s) \rangle] \\ &= \mathbb{E} [\langle \nabla f(x_t), \nabla f(x_s) \rangle] + \mathbb{E} [\mathbb{E} [\langle \nabla f_{\xi_s}(x_s) - \nabla f(x_s), \nabla f(x_t) \rangle \mid x_s]] \\ &\leq \mathbb{E} [\langle \nabla f(x_t), \nabla f(x_s) \rangle] + 3(t-s)\gamma L \rho^2 \end{aligned}$$

where the last inequality uses Lemma C.7 (with  $x_s$  regarded as initial point of the stochastic gradient descent). Now we apply this inequality and the induction hypothesis to (12):

$$\begin{aligned} & \mathbb{E} \left[ \|x_0 - x_{t+1}\|^2 \right] \\ &\leq \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^{t-1} \nabla f(x_s) \right\|^2 + t \rho^2 + (t-1)t(t+1)\gamma L \rho^2 \right. \\ &\quad \left. + \sum_{s=0}^{t-1} (2 \langle \nabla f(x_t), \nabla f(x_s) \rangle + 6(t-s)\gamma L \rho^2) + \|\nabla f(x_t)\|^2 + \rho^2 \right] \\ &= \gamma^2 (t \rho^2 + (t-1)t(t+1)\gamma L \rho^2 + 3t(t+1)\gamma L \rho^2 + \rho^2) \\ &\quad + \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^{t-1} \nabla f(x_s) \right\|^2 + 2 \left\langle \sum_{s=0}^{t-1} \nabla f(x_s), \nabla f(x_t) \right\rangle + \|\nabla f(x_t)\|^2 \right] \\ &= \gamma^2 (t+1)\rho^2 + t(t+1)(t+2)\gamma^3 L \rho^2 + \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^t \nabla f(x_s) \right\|^2 \right] \end{aligned}$$

where for the first equality we use  $\sum_{s=0}^{t-1} 6(t-s) = 3t(t+1)$ . This completes the induction.  $\square$

**Lemma C.9.** Let  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  be convex and  $L$ -smooth, and let  $x_0 \in \mathbb{R}^m$  be any initial point. Let  $x_1, \dots, x_t$  be generated by stochastic gradient descent

$$x_{s+1} = x_s - \gamma \nabla f_{\xi_s}(x_s)$$

with  $0 < \gamma \leq \frac{1}{L} \min \left\{ 1, \frac{1}{t-1} \right\}$ . Then

$$\mathbb{E} \left[ \|x_0 - x_t\|^2 \right] \leq \gamma^2 t^2 \|\nabla f(x_0)\|^2 + \gamma^2 t(1+2(t-1)(t+1)\gamma L)\rho^2.$$

*Proof.* Lemma C.8 gives

$$\mathbb{E} \left[ \|x_0 - x_t\|^2 \right] \leq \gamma^2 \mathbb{E} \left[ \left\| \sum_{s=0}^{t-1} \nabla f(x_s) \right\|^2 \right] + \gamma^2 t \rho^2 + (t-1)t(t+1)\gamma^3 L \rho^2. \quad (13)$$

Next, by Jensen's inequality and Lemma C.5,

$$\begin{aligned} \mathbb{E} \left[ \left\| \sum_{s=0}^{t-1} \nabla f(x_s) \right\|^2 \right] &\leq t \sum_{s=0}^{t-1} \mathbb{E} \left[ \|\nabla f(x_s)\|^2 \right] \\ &\leq t \sum_{s=0}^{t-1} \left( \|\nabla f(x_0)\|^2 + 2s\gamma L \rho^2 \right) \\ &\leq t^2 \|\nabla f(x_0)\|^2 + (t-1)t(t+1)\gamma L \rho^2 \end{aligned}$$

where the last inequality uses  $\sum_{s=0}^{t-1} 2s = t(t-1) \leq (t-1)(t+1)$ . Applying the above inequality to (13) we obtain the desired result.  $\square$

#### C.4 Proofs of Lemmas C.3 and C.2

*Proof of Lemma C.3.* Observe that given  $\mathbf{x}_{\tau p}$ , the sequence  $x_{\tau p}^i, \dots, x_{\tau(p+1)}^i$  is a sequence generated via stochastic gradient descent

$$x_{j+1}^i = x_j^i - \gamma \nabla f_{i,\xi_j^i}(x_j^i; x_{\tau p}^{-i})$$

for the  $L_i$ -smooth convex function  $f_i(\cdot; x_{\tau p}^{-i})$ , with  $x_{\tau p}^i$  as initial point, using the stochastic oracle  $\nabla f_{i,\xi^i}(\cdot; x_{\tau p}^{-i})$  satisfying (BV) (unbiased estimator of  $\nabla f_i(\cdot; x_{\tau p}^{-i})$  with variance at most  $\sigma_i^2$ ). Therefore, we can apply Lemma C.5 with

$$f(\cdot) = f_i(\cdot; x_{\tau p}^{-i}), \quad \rho^2 = \sigma_i^2, \quad x_0 = x_{\tau p}^i, \quad x_s = x_j^i, \quad L = L_i$$

and this immediately proves the desired statement. (Note that  $s$  is replaced with  $j - \tau p$  because  $x_j^i$  is obtained by  $j - \tau p$  steps of SGD from  $x_{\tau p}^i$ .)  $\square$

*Proof of Lemma C.2.* This is a direct consequence of Lemma C.9 with same choice of  $f, \rho^2, x_0$  and  $L$  as in the proof of Lemma C.3 and  $x_s = x_{\tau p+t}^i$ .  $\square$

### C.5 Remaining details in proof of Theorem 3.4

Note that the step-size condition of Lemma C.2 is satisfied by our step-size selection, as  $\gamma < \frac{2}{\ell\tau+2(\tau-1)L_{\max}\sqrt{\kappa}} \leq \frac{1}{L_{\max}(\tau-1)}$  (because  $\kappa \geq 1$ ). Now combine Lemmas C.1 and C.2 to obtain

$$\begin{aligned}
& \mathbb{E} \left[ \left\| \mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \right\|^2 \middle| \mathbf{x}_{\tau p} \right] \\
& \leq (1 + \alpha\gamma(\tau-1)) \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2 - 2\gamma\tau \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle \\
& \quad + \sum_{j=\tau p+1}^{\tau(p+1)-1} \sum_{i=1}^n \frac{\gamma L_i^2}{\alpha} \left( \gamma^2(j-\tau p)^2 \left\| \nabla f(x_{\tau p}^i; x_{\tau p}^{-i}) \right\|^2 + \gamma^2(j-\tau p)(1+2(j-\tau p-1)(j-\tau p+1)\gamma L_i)\sigma_i^2 \right) \\
& \quad + \sum_{i=1}^n \left( \gamma^2\tau^2 \left\| \nabla f(x_{\tau p}^i; x_{\tau p}^{-i}) \right\|^2 + \gamma^2\tau(1+2(\tau-1)(\tau+1)\gamma L_i)\sigma_i^2 \right) \\
& \leq (1 + \alpha\gamma(\tau-1)) \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2 - 2\gamma\tau \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle + \left( \gamma^2\tau^2 + \frac{\gamma^3 L_{\max}^2 \tau^2(\tau-1)}{3\alpha} \right) \left\| \mathbb{F}(\mathbf{x}_{\tau p}) \right\|^2 \\
& \quad + \gamma^2\tau \left( 1 + (\tau-1)\gamma L_{\max} \left( 2(\tau+1) + \frac{L_{\max}}{2\alpha} + \frac{\gamma L_{\max}^2}{2\alpha}(\tau+1)^2 \right) \right) \sigma^2
\end{aligned} \tag{14}$$

where for the last inequality, we replace all occurrences of  $L_i$ 's by  $L_{\max} = \max\{L_1, \dots, L_n\}$  and use the identities

$$\sigma^2 = \sum_{i=1}^n \sigma_i^2, \quad \left\| \mathbb{F}(\mathbf{x}_{\tau p}) \right\|^2 = \sum_{i=1}^n \left\| \nabla f_i(x_{\tau p}^i; x_{\tau p}^{-i}) \right\|^2$$

to eliminate the summations  $\sum_{i=1}^n$  and use the following elementary summation results:

$$\begin{aligned}
\sum_{j=\tau p+1}^{\tau(p+1)-1} (j-\tau p)^2 &= \frac{(\tau-1)\tau(2\tau-1)}{6} \leq \frac{(\tau-1)\tau^2}{3} \\
\sum_{j=\tau p+1}^{\tau(p+1)-1} (j-\tau p) &= \frac{(\tau-1)\tau}{2}
\end{aligned}$$

and

$$\sum_{j=\tau p+1}^{\tau(p+1)-1} (j-\tau p-1)(j-\tau p)(j-\tau p+1) = \frac{(\tau-2)(\tau-1)\tau(\tau+1)}{2} \leq \frac{(\tau-1)\tau(\tau+1)^2}{2}.$$

Now in (14), we use the assumption **(SCO)** to bound

$$\begin{aligned}
& -2\gamma\tau \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle + \left( \gamma^2\tau^2 + \frac{\gamma^3 L_{\max}^2 \tau^2(\tau-1)}{3\alpha} \right) \left\| \mathbb{F}(\mathbf{x}_{\tau p}) \right\|^2 \\
& \leq - \left( 2\gamma\tau - \ell \left( \gamma^2\tau^2 + \frac{\gamma^3 L_{\max}^2 \tau^2(\tau-1)}{3\alpha} \right) \right) \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle \\
& = -\gamma\tau \left( 2 - \gamma\ell\tau - \frac{\gamma^2\ell L_{\max}^2 \tau(\tau-1)}{3\alpha} \right) \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle.
\end{aligned} \tag{15}$$

Provided that

$$2 - \gamma\ell\tau - \frac{\gamma^2\ell L_{\max}^2 \tau(\tau-1)}{3\alpha} \geq 0, \tag{16}$$

we can again upper bound (15) using the assumption **(QSM)**:

$$\begin{aligned}
& -\gamma\tau \left( 2 - \gamma\ell\tau - \frac{\gamma^2\ell L_{\max}^2 \tau(\tau-1)}{3\alpha} \right) \langle \mathbf{x}_{\tau p} - \mathbf{x}_*, \mathbb{F}(\mathbf{x}_{\tau p}) \rangle \\
& \leq -\gamma\tau \left( 2 - \gamma\ell\tau - \frac{\gamma^2\ell L_{\max}^2 \tau(\tau-1)}{3\alpha} \right) \mu \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2.
\end{aligned}$$

We plug this into (14) and rearrange the terms to obtain

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \right\|^2 \middle| \mathbf{x}_{\tau p} \right] \\ & \leq \left( 1 + \alpha \gamma (\tau - 1) - \gamma \tau \left( 2 - \gamma \ell \tau - \frac{\gamma^2 \ell L_{\max}^2 \tau (\tau - 1)}{3\alpha} \right) \mu \right) \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2 \\ & \quad + \gamma^2 \tau \left( 1 + (\tau - 1) \gamma L_{\max} \left( 2(\tau + 1) + \frac{L_{\max}}{2\alpha} + \frac{\gamma L_{\max}^2}{2\alpha} (\tau + 1)^2 \right) \right) \sigma^2. \end{aligned} \quad (17)$$

Now, we optimize the coefficient of the  $\left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2$  term in (17) by taking

$$\alpha = \underset{\alpha > 0}{\operatorname{argmin}} \alpha \gamma (\tau - 1) + \frac{\gamma^3 \ell L_{\max}^2 \tau^2 (\tau - 1) \mu}{3\alpha} = \gamma \tau L_{\max} \sqrt{\frac{\ell \mu}{3}}.$$

With this choice of  $\alpha$ , the bound (17) becomes

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \right\|^2 \middle| \mathbf{x}_{\tau p} \right] \\ & \leq \left( 1 - \gamma \tau \mu \left( 2 - \gamma \ell \tau - 2(\tau - 1) \gamma L_{\max} \sqrt{\frac{\ell}{3\mu}} \right) \right) \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2 \\ & \quad + \gamma^2 \tau \left( 1 + (\tau - 1) \gamma L_{\max} \left( 2(\tau + 1) + \frac{1}{2\gamma \tau \sqrt{\ell \mu / 3}} + \frac{L_{\max} (\tau + 1)^2}{2\tau \sqrt{\ell \mu / 3}} \right) \right) \sigma^2 \\ & \leq (1 - \gamma \tau \mu \zeta) \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2 + \gamma^2 \tau \sigma^2 \left( 1 + (\tau - 1) \left( 4\gamma \tau L_{\max} + \frac{L_{\max}}{2\tau \sqrt{\ell \mu / 3}} + \frac{\gamma \tau L_{\max}^2}{\sqrt{\ell \mu / 3}} \right) \right) \end{aligned} \quad (18)$$

where for the last inequality, we use  $\tau + 1 \leq 2\tau$  and make the substitution

$$\zeta = 2 - \gamma \ell \tau - 2(\tau - 1) \gamma L_{\max} \sqrt{\frac{\ell}{3\mu}} = 2 - \gamma \ell \tau - 2(\tau - 1) \gamma L_{\max} \sqrt{\kappa/3}.$$

Note that with our choice  $\alpha = \gamma \tau L_{\max} \sqrt{\frac{\ell \mu}{3}}$  and  $0 < \gamma < \frac{2}{\ell \tau + 2(\tau - 1) L_{\max} \sqrt{\kappa}}$ , the condition (16) is satisfied because

$$\begin{aligned} 2 - \gamma \ell \tau - \frac{\gamma^2 \ell L_{\max}^2 \tau (\tau - 1)}{3\alpha} & \geq 2 - \gamma \ell \tau - \frac{\gamma^2 \ell L_{\max}^2 \tau (\tau - 1)}{3\alpha} \\ & = 2 - \gamma \ell \tau - (\tau - 1) \gamma L_{\max} \sqrt{\frac{\ell}{3\mu}} \\ & \geq 2 - \gamma (\ell \tau + (\tau - 1) L_{\max} \sqrt{\kappa}) > 0. \end{aligned}$$

Finally, unrolling the recursion (18) using the following simple lemma, with  $a_p = \mathbb{E} \left[ \left\| \mathbf{x}_{\tau p} - \mathbf{x}_* \right\|^2 \right]$ ,  $A = \tau \mu \zeta$  and

$$B = \tau \sigma^2 \left( 1 + (\tau - 1) \left( 4\gamma \tau L_{\max} + \frac{L_{\max}}{2\tau \sqrt{\ell \mu / 3}} + \frac{\gamma \tau L_{\max}^2}{\sqrt{\ell \mu / 3}} \right) \right)$$

gives the desired rate. (Note that  $\gamma A = \gamma \tau \mu \zeta \leq \gamma \tau \mu (2 - \gamma \ell \tau) \leq \gamma \ell \tau (2 - \gamma \ell \tau) \leq 1$ .)

**Lemma C.10.** Let  $\gamma, A, B > 0$  with  $\gamma A \leq 1$ . If a sequence  $a_0, \dots, a_R \in \mathbb{R}$  satisfies

$$a_{p+1} \leq (1 - \gamma A) a_p + \gamma^2 B$$

for  $p = 0, \dots, R - 1$ , then  $a_R \leq (1 - \gamma A)^R a_0 + \frac{\gamma B}{A}$ .

*Proof of Lemma C.10.* As there is nothing to prove if  $\gamma A = 1$ , suppose  $\gamma A < 1$ . Recursively applying the given inequality we have

$$a_R \leq (1 - \gamma A)a_{R-1} + \gamma^2 B \leq \dots \leq (1 - \gamma A)^R a_0 + \gamma^2 B \sum_{p=0}^{R-1} (1 - \gamma A)^p.$$

Now apply the bound  $\sum_{p=0}^{R-1} (1 - \gamma A)^p \leq \sum_{p=0}^{\infty} (1 - \gamma A)^p = \frac{1}{1 - (1 - \gamma A)} = \frac{1}{\gamma A}$  to the above inequality.  $\square$

## C.6 Proof of Corollary 3.5

First, because  $\eta > \kappa\tau$ , we have

$$\gamma < \frac{1}{\mu\kappa\tau \left(1 + \frac{2L_{\max}}{\sqrt{\ell\mu}}\right)} = \frac{1}{\ell\tau \left(1 + \frac{2L_{\max}}{\sqrt{\ell\mu}}\right)} \leq \frac{1}{\ell\tau + 2(\tau - 1)L_{\max}\sqrt{\frac{\ell}{\mu}}} = \frac{1}{\ell\tau + 2(\tau - 1)L_{\max}\sqrt{\kappa}}.$$

Hence we can apply Theorem 3.4. Now observe that  $\zeta > 2 - \gamma(\ell\tau + 2(\tau - 1)L_{\max}\sqrt{\kappa}) > 1$ , and  $(1 - u)^R \leq e^{-uR}$  for  $u < 1$ , so

$$(1 - \gamma\tau\mu\zeta)^R \leq e^{-\gamma\mu\zeta\tau R} \leq e^{-\gamma\mu T} = e^{-2\log\eta} = \frac{1}{\eta^2} = \frac{4(\log\eta)^2(1 + 2q)^2}{T^2} = \tilde{\mathcal{O}}\left(\frac{(1 + q)^2}{T^2}\right)$$

where we use  $T = 2(1 + 2q)\eta\log\eta$  and remove the factor  $\log\eta < \log T$  within the  $\tilde{\mathcal{O}}$  notation. Next, for the terms proportional to  $\sigma^2$ , we have

$$\begin{aligned} & \left(1 + (\tau - 1) \left(4\gamma\tau L_{\max} + \frac{L_{\max}}{2\tau\sqrt{\ell\mu/3}} + \frac{\gamma\tau L_{\max}^2}{\sqrt{\ell\mu/3}}\right)\right) \frac{\gamma\sigma^2}{\mu\zeta} \\ & \leq \frac{\gamma\sigma^2}{\mu} \left(1 + \tau \left(4\gamma\tau L_{\max} + \frac{\sqrt{3}q}{2\tau} + \sqrt{3}\gamma\tau L_{\max}q\right)\right) \\ & \leq \frac{\gamma\sigma^2}{\mu} \left(1 + \frac{\sqrt{3}q}{2}\right) + \frac{\gamma^2\tau^2 L_{\max}\sigma^2}{\mu} (4 + \sqrt{3}q) \\ & = \frac{\sigma^2(1 + \sqrt{3}q/2)}{\mu^2\eta(1 + 2q)} + \frac{\tau^2 L_{\max}\sigma^2(4 + \sqrt{3}q)}{\mu^3\eta^2(1 + 2q)^2} \\ & = \tilde{\mathcal{O}}\left(\frac{(1 + q)\sigma^2}{\mu^2 T} + \frac{(1 + q)\tau^2 L_{\max}\sigma^2}{\mu^3 T^2}\right). \end{aligned}$$

Combining these with Theorem 3.4 we arrive at the desired conclusion.

## C.7 Proof of Theorem 3.6

Note that we use constant step-size  $\gamma_k \equiv \gamma_{\tau p}$  within each communication round  $p$ , i.e., for  $\tau p \leq k \leq \tau(p + 1) - 1$ , so we can apply the bound (18) from the proof of Theorem 3.4, provided that

$$\gamma_{\tau p} \leq \frac{1}{\ell\tau + 2(\tau - 1)L_{\max}\sqrt{\kappa}}.$$

This clearly holds true when  $p < 2(1 + 2q)\kappa - 1$ , and when  $p \geq 2(1 + 2q)\kappa - 1$  then

$$\gamma_{\tau p} = \frac{1}{\tau\mu} \frac{2p + 1}{(p + 1)^2} < \frac{1}{\tau\mu} \frac{2}{p + 1} \leq \frac{1}{\tau\mu} \frac{1}{(1 + 2q)\kappa} = \frac{1}{\ell\tau + 2\tau L_{\max}\sqrt{\kappa}}$$

so we see that the step-size condition is satisfied. Furthermore we have

$$\zeta_{\tau p} = 2 - \gamma_{\tau p}\ell\tau - 2(\tau - 1)\gamma_{\tau p}L_{\max}\sqrt{\kappa/3} > 1,$$

so (18), with  $q = \frac{L_{\max}}{\sqrt{\ell\mu}}$  and taking expectation with respect to  $\mathbf{x}_{\tau p}$ , gives

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \|^2 \right] &\leq (1 - \gamma_{\tau p} \tau \mu \zeta_{\tau p}) \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_* \|^2 \right] \\ &\quad + \gamma_{\tau p}^2 \tau \sigma^2 \left( 1 + (\tau - 1) \left( \gamma_{\tau p} \tau L_{\max} (4 + \sqrt{3}q) + \frac{\sqrt{3}}{2\tau q} \right) \right) \\ &\leq (1 - \gamma_{\tau p} \tau \mu) \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_* \|^2 \right] + (1 + q) \gamma_{\tau p}^2 \tau \sigma^2 + 4(1 + q) \gamma_{\tau p}^3 \tau^2 (\tau - 1) L_{\max} \sigma^2. \end{aligned} \quad (19)$$

For  $p \geq 2(1 + 2q)\kappa - 1$ , plugging in  $\gamma_{\tau p} = \frac{1}{\tau\mu} \frac{2p+1}{(p+1)^2}$  we obtain

$$\mathbb{E} \left[ \|\mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \|^2 \right] \leq \frac{p^2}{(p+1)^2} \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_* \|^2 \right] + \frac{(2p+1)^2 \sigma^2 (1+q)}{\tau\mu^2 (p+1)^4} \left( 1 + \frac{4(\tau-1)L_{\max}(2p+1)}{\mu(p+1)^2} \right).$$

Multiplying  $\tau^2(p+1)^2$  to both sides and upper-bounding  $\frac{2p+1}{p+1} \leq 2$ , we obtain

$$(\tau(p+1))^2 \mathbb{E} \left[ \|\mathbf{x}_{\tau(p+1)} - \mathbf{x}_* \|^2 \right] \leq (\tau p)^2 \mathbb{E} \left[ \|\mathbf{x}_{\tau p} - \mathbf{x}_* \|^2 \right] + \frac{4(1+q)\tau\sigma^2}{\mu^2} \left( 1 + \frac{8(\tau-1)L_{\max}}{\mu(p+1)} \right).$$

Let  $p_0 = \lceil 2(1 + 2q)\kappa - 1 \rceil$ . Chaining the above inequality for  $p = p_0, \dots, R-1$  gives

$$\begin{aligned} &(\tau R)^2 \mathbb{E} \left[ \|\mathbf{x}_{\tau R} - \mathbf{x}_* \|^2 \right] \\ &\leq (\tau p_0)^2 \mathbb{E} \left[ \|\mathbf{x}_{\tau p_0} - \mathbf{x}_* \|^2 \right] + \frac{4(1+q)\tau(R-p_0)\sigma^2}{\mu^2} + \frac{32(1+q)\tau(\tau-1)L_{\max}\sigma^2}{\mu^3} \sum_{p=p_0}^{R-1} \frac{1}{p+1} \\ &\leq (\tau p_0)^2 \mathbb{E} \left[ \|\mathbf{x}_{\tau p_0} - \mathbf{x}_* \|^2 \right] + \frac{4(1+q)\tau(R-p_0)\sigma^2}{\mu^2} + \frac{32(1+q)\tau^2 L_{\max} \sigma^2 \log(R/p_0)}{\mu^3} \end{aligned}$$

where we use  $\sum_{p=p_0}^{R-1} \frac{1}{p+1} \leq \int_{p_0}^R \frac{dp}{p} = \log \frac{R}{p_0}$ . Now substitute  $T = \tau R$  using the upper bounds  $\tau(R-p_0) \leq \tau R = T$  and  $\log(R/p_0) \leq \log T$ , we can write

$$T^2 \mathbb{E} \left[ \|\mathbf{x}_T - \mathbf{x}_* \|^2 \right] \leq (\tau p_0)^2 \mathbb{E} \left[ \|\mathbf{x}_{\tau p_0} - \mathbf{x}_* \|^2 \right] + \frac{4(1+q)T\sigma^2}{\mu^2} + \frac{32(1+q)\tau^2 L_{\max} \sigma^2 \log T}{\mu^3}. \quad (20)$$

As  $\gamma_k$  is constantly  $\gamma_0 = \frac{1}{\ell\tau(1+2q)}$  over rounds  $p = 0, \dots, p_0 - 1$ , we can directly apply Theorem 3.4 with  $R = p_0$  and similar simplification of the  $\sigma^2$ -terms as in (19) to bound

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{x}_{\tau p_0} - \mathbf{x}_* \|^2 \right] &\leq \left( 1 - \frac{\mu}{\ell(1+2q)} \right)^{p_0} \|\mathbf{x}_0 - \mathbf{x}_* \|^2 + \frac{(1+q)\gamma_0\sigma^2}{\mu} (1 + 4\gamma_0\tau(\tau-1)L_{\max}) \\ &\leq \left( 1 - \frac{1}{\kappa(1+2q)} \right)^{\kappa(1+2q)} \|\mathbf{x}_0 - \mathbf{x}_* \|^2 + \frac{\sigma^2}{\ell\mu\tau} \left( 1 + \frac{4(\tau-1)L_{\max}}{\ell(1+2q)} \right) \\ &\leq \frac{\|\mathbf{x}_0 - \mathbf{x}_* \|^2}{e} + \frac{\sigma^2}{\ell\mu\tau} \left( 1 + \frac{2\tau}{\sqrt{\kappa}} \right), \end{aligned}$$

where the second line uses  $p_0 \geq 2(1 + 2q)\kappa - 1 \geq \kappa(1 + 2q)$ , and the third line uses the bound  $(1 - \frac{1}{t})^t \leq \frac{1}{e}$  for  $t > 1$  and  $\frac{4(\tau-1)L_{\max}}{\ell(1+2q)} \leq \frac{4q\tau\sqrt{\ell\mu}}{\ell(1+2q)} \leq 2\tau\sqrt{\frac{\ell}{\mu}} = \frac{2\tau}{\sqrt{\kappa}}$ . Now plugging this into (20) and dividing both sides by  $T^2$  we obtain

$$\begin{aligned} &\mathbb{E} \left[ \|\mathbf{x}_T - \mathbf{x}_* \|^2 \right] \\ &\leq \frac{p_0^2 \tau^2 \|\mathbf{x}_0 - \mathbf{x}_* \|^2}{eT^2} + \frac{\tau p_0^2 \sigma^2}{\ell\mu T^2} \left( 1 + \frac{2\tau}{\sqrt{\kappa}} \right) + \frac{4(1+q)\sigma^2}{\mu^2 T} + \frac{32(1+q)\tau^2 L_{\max} \sigma^2 \log T}{\mu^3 T^2} \\ &\leq \frac{4(1+2q)^2 \kappa^2 \tau^2 \|\mathbf{x}_0 - \mathbf{x}_* \|^2}{eT^2} + \frac{4(1+q)\sigma^2}{\mu^2 T} + \frac{4(1+2q)^2 \kappa \tau \sigma^2}{\mu^2 T^2} \left( 1 + \frac{2\tau}{\sqrt{\kappa}} \right) + \frac{32(1+q)\tau^2 L_{\max} \sigma^2 \log T}{\mu^3 T^2}. \end{aligned}$$

which is the desired result.

## D Details of Numerical Experiments

Experiments were conducted using a personal MacBook with an Apple M3 chip and 16GB RAM.

### D.1 Quadratic $n$ -player game

We set  $n = 5$ ,  $d = 10$  and  $M = 100$ . The matrices  $\mathbf{A}_{i,m}$  are generated randomly with their eigenvalues in the range  $[\mu_{\mathbf{A}}, L_{\mathbf{A}}]$  ( $0 < \mu_{\mathbf{A}} < L_{\mathbf{A}}$ ). Similarly, for  $1 \leq i < j \leq n$ , we generate the matrices  $\mathbf{B}_{i,j,m}$  randomly with their eigenvalues in  $[0, L_{\mathbf{B}}]$ . Notably, we set  $\mathbf{B}_{j,i,m} = -\mathbf{B}_{i,j,m}^T$  for  $1 \leq j < i \leq n$ . With this condition, we can ensure that the  $n$ -player game (2) satisfies the (QSM) assumption, regardless of the values of  $\mu_{\mathbf{A}}$ ,  $L_{\mathbf{A}}$  and  $L_{\mathbf{B}}$ . We show below why this is the case.

Recall that we have

$$f_i(x^1, \dots, x^n) = \frac{1}{2} \langle x^i, \mathbf{A}_i x^i \rangle + \langle a_i, x^i \rangle + \sum_{j \neq i} \langle x^i, \mathbf{B}_{i,j} x^j \rangle$$

for  $i = 1, \dots, n$ . Differentiating  $f_i$  with respect to  $x^i$ , we get

$$\nabla f_i(x^i; x^{-i}) = \mathbf{A}_i x^i + a_i + \sum_{j \neq i} \mathbf{B}_{i,j} x^j$$

and thus

$$\begin{aligned} \nabla f_i(x^i; x^{-i}) - \nabla f_i(x_{\star}^i; x_{\star}^{-i}) &= \left( \mathbf{A}_i x^i + a_i + \sum_{j \neq i} \mathbf{B}_{i,j} x^j \right) - \left( \mathbf{A}_i x_{\star}^i + a_i + \sum_{j \neq i} \mathbf{B}_{i,j} x_{\star}^j \right) \\ &= \mathbf{A}_i(x^i - x_{\star}^i) + \sum_{j \neq i} \mathbf{B}_{i,j}(x^j - x_{\star}^j) \end{aligned}$$

and

$$\begin{aligned} \langle \mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{x}_{\star}), \mathbf{x} - \mathbf{x}_{\star} \rangle &= \sum_{i=1}^n \langle \nabla f_i(x^i; x^{-i}) - \nabla f_i(x_{\star}^i; x_{\star}^{-i}), x^i - x_{\star}^i \rangle \\ &= \sum_{i=1}^n \langle x^i - x_{\star}^i, \mathbf{A}_i(x^i - x_{\star}^i) \rangle + \sum_{i=1}^n \sum_{j \neq i} \langle x^i - x_{\star}^i, \mathbf{B}_{i,j}(x^j - x_{\star}^j) \rangle. \end{aligned}$$

Now, the double summation term vanishes because for any  $i \neq j$ ,

$$\langle x^i - x_{\star}^i, \mathbf{B}_{i,j}(x^j - x_{\star}^j) \rangle + \langle x^j - x_{\star}^j, \mathbf{B}_{j,i}(x^i - x_{\star}^i) \rangle = 0$$

due to the condition  $\mathbf{B}_{j,i} = -\mathbf{B}_{i,j}^T$ . Therefore, provided that each  $\mathbf{A}_i \succeq \mu I$  we see that  $\mathbb{F}$  satisfies (QSM) (in fact, the same argument with arbitrary  $\mathbf{y}$  in place of  $\mathbf{x}_{\star}$  shows that  $\mathbb{F}$  is  $\mu$ -strongly monotone).

### D.2 Distributed mobile robot control

We follow the same choice of parameter values  $a_i, b_i, x_{\text{anc}}^i, h_{ij}$  within (3) from [60]:  $n = 5$ ,  $d = 1$ ,  $a_i = 10 + i/6$ ,  $b_i = i/6$ ,

$$(x_{\text{anc}}^1, x_{\text{anc}}^2, x_{\text{anc}}^3, x_{\text{anc}}^4, x_{\text{anc}}^5) = (1, -4, 8, -9, 13)$$

and

$$(h_{ij})_{\substack{1 \leq i \leq 5 \\ 1 \leq j \leq 5}} = \begin{bmatrix} 0 & 5 & -7 & 9 & -8 \\ -5 & 0 & -6 & 2 & -9 \\ 7 & 6 & 0 & 7 & -4 \\ -9 & -2 & -7 & 0 & -2 \\ 8 & 9 & 4 & 2 & 0 \end{bmatrix}.$$

We add Gaussian noise with  $\sigma^2 = 100$  to the gradients to simulate stochasticity. In this setup, all our theoretical assumptions are satisfied [60].

## E Additional Experiments

### E.1 Quadratic $n$ -player game with step-size tuning

In this experiment, we simulate the scenario where we do not know the precise theoretical parameters in advance. For each  $\tau \in \{1, 2, 4, 5, 8, 20\}$ , we tune  $\gamma$  by running **PEARL-SGD** with each  $\gamma \in \{10^{-1}, 10^{-2}, \dots, 10^{-6}\}$ , and plot the best relative error  $\|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 / \|\mathbf{x}_0 - \mathbf{x}_*\|^2$  (y-axis) versus the communication round index  $p$  (x-axis). Figure 5a presents results from Deterministic **PEARL-SGD**, and Figure 5b presents results under stochasticity, imposed by mini-batching from the finite sum. The results demonstrate that in practice, we can use  $(\tau, \gamma)$  as tunable hyperparameters to achieve the best communication complexity.

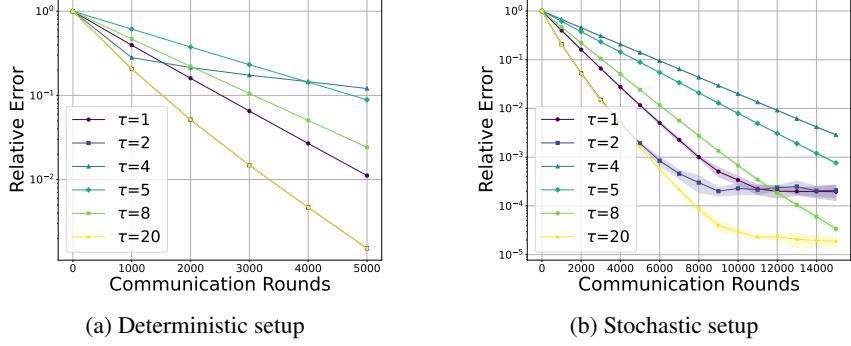


Figure 5: Performance plots for **PEARL-SGD** on the  $n$ -player game (2) with different values of  $\tau$ . For each  $\tau$ , we use the empirically tuned step-size  $\gamma \in \{10^{-1}, 10^{-2}, \dots, 10^{-6}\}$  for the best relative error  $\|\mathbf{x}_{\tau p} - \mathbf{x}_*\|^2 / \|\mathbf{x}_0 - \mathbf{x}_*\|^2$ . Figure 5a shows the result from deterministic setup and 5b shows the stochastic setup.

### E.2 Distributed mobile robot control

In Fig. 6, we plot the local objective values  $f_i$  in the setup (3) obtained from **PEARL-SGD** for each robot (player)  $i = 1, \dots, 5$  in the case  $\tau = 5$ . Generally in games, the objectives  $f_i$  can have both cooperative and competitive components. After the cooperative components in each  $f_i$  are sufficiently reduced,  $f_i$ 's can oscillate due to the competing interests of players until an equilibrium is found, and then finally stabilize around a certain level.

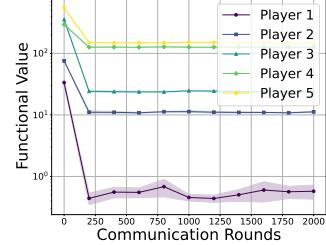


Figure 6: Local objective  $f_i$  in mobile robot control setup.

## F Discussion on Theoretical Assumptions

### F.1 Possible simplification of assumptions: Assuming cocoercivity of $\mathbb{F}$

In fact, the convergence of **PEARL-SGD** can still be proved even if the three assumptions **(CVX)**, **(SM)** and **(SCO)** are replaced with the single assumption that  $\mathbb{F}: \mathbb{R}^D \rightarrow \mathbb{R}^D$  is  $\frac{1}{\ell}$ -cocoercive, i.e.,

$$\langle \mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{\ell} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^D. \quad (\text{COCO})$$

In the subsequent paragraphs, we explain in detail why this is the case. However, we emphasize here that if we derived all convergence theory using **(COCO)** in place of **(CVX)**, **(SM)** and **(SCO)** and did not distinguish the role of  $L_i$ 's (the local Lipschitzness parameters from **(SM)**) from that of  $\ell$ , then the resulting convergence rates would have become much more pessimistic (worse) in many cases. Therefore, in our work, we choose to use the current set of assumptions. It allows us to more clearly present the tight dependency of convergence rates to  $L_i$ 's. Also note that assuming **(CVX)**, **(SM)** and **(SCO)** is strictly more general than assuming **(COCO)**, as we illustrate in Appendix F.2.

**(COCO)** implies **(CVX)**, **(SM)** and **(SCO)**. Trivially, **(COCO)** implies **(SCO)**. Furthermore, if  $\mathbb{F}$  is  $\frac{1}{\ell}$ -cocoercive, then  $\mathbb{F}$  is monotone:

$$\langle \mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^D, \quad (21)$$

and  $\ell$ -Lipschitz continuous:

$$\|\mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{y})\| \leq \ell \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^D. \quad (22)$$

In particular, for each  $i = 1, \dots, n$ , we can take

$$\mathbf{x} = (x^1, \dots, x^{i-1}, x^i, x^{i+1}, \dots, x^n), \quad \mathbf{y} = (x^1, \dots, x^{i-1}, y^i, x^{i+1}, \dots, x^n) \quad (23)$$

in (21), which gives

$$\langle \nabla f_i(x^i; x^{-i}) - \nabla f_i(y^i; x^{-i}), x^i - y^i \rangle \geq 0$$

for any  $x^i, y^i \in \mathbb{R}^{d_i}$  and  $x^{-i} \in \mathbb{R}^{D-d_i}$ . That is, the gradient of  $f_i(\cdot; x^{-i}) : \mathbb{R}^{d_i} \rightarrow \mathbb{R}$  is a monotone operator on  $\mathbb{R}^{d_i}$ , and this implies that  $f_i(\cdot; x^{-i})$  is convex, i.e., **(CVX)** holds. Similarly, plugging the choice (23) into (22) we obtain

$$\|\nabla f_i(x^i; x^{-i}) - \nabla f_i(y^i; x^{-i})\| \leq \ell \|x^i - y^i\|,$$

showing that **(SM)** holds, with  $L_i = \ell$ . Therefore, all theorems from the main paper hold under the assumptions **(QSM)**, **(COCO)**, and **(BV)**, with  $\ell$  in place of  $L_{\max}$  in step-size restrictions and convergence rates.

**What do we lose by replacing  $L_{\max}$  with  $\ell$ ?** The previous discussion shows that we can assume **(COCO)** and replace all occurrences of  $L_{\max}$  with  $\ell$  within the theory. In this case, however, the step-size conditions in Theorems 3.3 and 3.4 become

$$\gamma \leq \frac{1}{\ell(\tau + 2(\tau - 1)\sqrt{\kappa})} = \mathcal{O}\left(\frac{1}{\ell\tau\sqrt{\kappa}}\right), \quad (24)$$

and the  $\sqrt{\kappa}$  factor in the denominator is undesirable as it significantly restricts the range of step-size one can use if  $\kappa$  is large. Furthermore, in Corollary 3.5 and Theorem 3.6, the factor  $q$  becomes  $\sqrt{\frac{\ell}{\mu}} = \sqrt{\kappa}$ , causing the constant factors in the convergence bounds to potentially become large.

However, there are many cases where  $L_{\max} \ll \ell$ , showing why it is beneficial to keep the dependency on  $L_{\max}$  tight as we do. As an abstract example, when  $\mathbb{F}$  is a generic  $\mu$ -strongly monotone and  $M$ -Lipschitz continuous operator, the tight (smallest) cocoercivity parameter one can guarantee on  $\mathbb{F}$  is  $\ell = M^2/\mu$  [33] (tightness can be shown using, e.g., the scaled relative graph theory in [115], [114, Chapter 13]). On the other hand, we have

$$L_{\max} \leq \max_{i=1, \dots, n} \sup_{\substack{\mathbf{x}=(x^i, x^{-i}), \mathbf{y}=(y^i, x^{-i}) \\ x^i \neq y^i}} \frac{\|\mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} \leq \sup_{\mathbf{x} \neq \mathbf{y}} \frac{\|\mathbb{F}(\mathbf{x}) - \mathbb{F}(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} = M,$$

i.e.,  $M$  is an upper bound on  $L_{\max}$  (better than  $\ell$ ). Therefore,  $\ell$  is at least  $\frac{\ell}{M} = \frac{\ell}{\sqrt{\ell\mu}} = \sqrt{\kappa}$  times larger than  $L_{\max}$ , and the largest step-size allowed in Theorems 3.3 and 3.4 is

$$\frac{1}{\ell\tau + 2(\tau-1)L_{\max}\sqrt{\kappa}} = \Omega\left(\frac{1}{\ell\tau}\right)$$

which is in contrast with (24) where we used  $\ell$  in place of  $L_{\max}$  and obtained  $\sqrt{\kappa}$  times smaller step-size range. Additionally, note that in this case  $q = \frac{L_{\max}}{\sqrt{\ell\mu}} = \frac{L_{\max}}{M} \leq 1$  in Corollary 3.5 and Theorem 3.6, so we can avoid the  $\kappa$ -dependent factors appearing in the convergence results.

We demonstrate another problem class for which  $L_{\max} \ll \ell$ . Consider a two-player matrix game, regularized by adding strongly convex (resp. strongly concave) quadratic terms in  $x$  (resp.  $y$ ):

$$\underset{u \in \mathbb{R}^m}{\text{minimize}} \underset{v \in \mathbb{R}^m}{\text{maximize}} \mathcal{L}(u, v) = \frac{\mu}{2} \|u\|^2 + g^\top u + u^\top \mathbf{B}v - h^\top v - \frac{\mu}{2} \|v\|^2 \quad (25)$$

where  $\mathbf{B} \in \mathbb{R}^{m \times m}$ ,  $g, h \in \mathbb{R}^m$ . In our  $n$ -player game notation, the first and second players respectively use the objective function  $f_1(x^1; x^2) = \mathcal{L}(x^1, x^2)$  and  $f_2(x^2; x^1) = -\mathcal{L}(x^1, x^2)$ . In this case, the operator  $\mathbf{F}$  is  $\mu$ -strongly monotone with  $\mu$  and  $M$ -Lipschitz continuous with parameter  $M \geq \sqrt{\|\mathbf{B}\|_2^2 + \mu^2} \geq \|\mathbf{B}\|_2$ . Note that the cocoercivity parameter  $\ell$  is at least  $M$  (and at most  $M^2/\mu$ ). On the other hand,

$$\nabla f_1(x^1; x^2) = \mu x^1 + g + \mathbf{B}x^2, \quad \nabla f_2(x^2; x^1) = \mu x^2 + h - \mathbf{B}^\top x^1,$$

so the Lipschitz constant for  $\nabla f_1$  with  $x^2$  fixed (resp.  $\nabla f_2$  with  $x^1$  fixed) is  $\mu$ , i.e.,  $L_{\max} = \mu$ . Therefore, we have  $L_{\max} \ll \ell$  in this scenario, as strength of regularization  $\mu$  is usually small compared to the smoothness parameter  $M$ . The same principle applies to the  $n$ -player analogue of this setup we use in Section 4.1, where each player has the objective function

$$f_i(x^i; x^{-i}) = \frac{1}{2} \langle x^i, \mathbf{A}_i x^i \rangle + \langle a_i, x^i \rangle + \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \langle x^i, \mathbf{B}_{i,j} x^j \rangle$$

with  $\mathbf{B}_{j,i} = -\mathbf{B}_{i,j}^\top$ . If the quadratic terms are the small regularization terms introduced to induce convergence, so that  $\mathbf{A}_i = \mu \mathbf{I}$  with  $\mu \ll \|\mathbf{B}_{i,j}\|_2$ , then we have  $L_{\max} = \mu \ll \max_{i \neq j} \|\mathbf{B}_{i,j}\|_2 \leq \ell$ .

## F.2 Example of non-cocoercive $\mathbf{F}$ satisfying (CVX), (SM), (QSM) and (SCO)

Consider the two-player game where two players have the objectives

$$\begin{aligned} f_1(u; v) &= \frac{u^2}{2} \varphi(v) \\ f_2(v; u) &= \frac{v^2}{2} \varphi(u) \end{aligned}$$

where  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  is defined by

$$\varphi(t) = (\mu + (\ell - \mu) \sin^2 t).$$

Here  $0 < \mu < \ell$ , and we use the notation  $\mathbf{x} = (u, v) \in \mathbb{R} \times \mathbb{R}$  instead of  $\mathbf{x} = (x^1, x^2)$  for better readability. Note that because  $\varphi$  satisfies

$$0 < \mu \leq \varphi(t) \leq \ell, \quad \forall t \in \mathbb{R},$$

$f_1(\cdot, v): \mathbb{R} \rightarrow \mathbb{R}$  is convex (quadratic) for any  $v \in \mathbb{R}$ , and so is  $f_2(u, \cdot)$  for any  $u \in \mathbb{R}$ . Therefore, this game satisfies (CVX). For any  $\mathbf{x} = (u, v)$ , we have

$$\mathbf{F}(\mathbf{x}) = (\nabla_u f_1(u; v), \nabla_v f_2(v; u)) = (u\varphi(v), v\varphi(u)).$$

Therefore, the unique equilibrium of the game is  $\mathbf{x}_* = (u_*, v_*) = (0, 0)$ . Additionally, observe that

$$\nabla_{uu} f_1(u; v) = \varphi(v) \in [\mu, \ell], \quad \nabla_{vv} f_2(v; u) = \varphi(u) \in [\mu, \ell].$$

In particular, the both second derivatives are bounded, so (SM) is satisfied. Next, we have

$$\langle \mathbf{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle = u^2 \varphi(v) + v^2 \varphi(u) \geq \mu(u^2 + v^2) = \mu \|\mathbf{x} - \mathbf{x}_*\|^2,$$

i.e.,  $\mathbb{F}$  satisfies **(QSM)**. Finally, we have

$$\|\mathbb{F}(\mathbf{x})\|^2 = u^2\varphi(v)^2 + v^2\varphi(u)^2 \leq \max\{\varphi(v), \varphi(u)\} (u^2\varphi(v) + v^2\varphi(u)) \leq \ell \langle \mathbb{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle,$$

showing that  $\mathbb{F}$  satisfies **(SCO)**.

On the other hand,  $\mathbb{F}$  is not cocoercive with respect to any parameter; in fact, it is not even Lipschitz continuous nor monotone. Observe that the cross-derivatives

$$\nabla_{uv} f_1(u; v) = (\ell - \mu)u \sin(2v), \quad \nabla_{vuu} f_2(u; v) = (\ell - \mu)v \sin(2u)$$

are unbounded over  $\mathbb{R} \times \mathbb{R}$ , so  $\mathbb{F}$  is cannot be Lipschitz continuous with any fixed parameter. Furthermore, we have

$$(D\mathbb{F} + D\mathbb{F}^\top)(u, v) = \begin{bmatrix} 2\varphi(v) & (\ell - \mu)(u \sin(2v) + v \sin(2u)) \\ (\ell - \mu)(u \sin(2v) + v \sin(2u)) & 2\varphi(u) \end{bmatrix}$$

so with  $u = v = (2N + \frac{1}{2})\pi$ , we have

$$\begin{aligned} \det(D\mathbb{F} + D\mathbb{F}^\top)(u, v) &= 4\varphi^2 \left( \left(2N + \frac{1}{2}\right)\pi \right) - 4(\ell - \mu)^2 \left(2N + \frac{1}{2}\right)^2 \pi^2 \\ &= 4\ell^2 - 4(\ell - \mu)^2 \left(2N + \frac{1}{2}\right)^2 \pi^2 \\ &< 0 \end{aligned}$$

provided that  $N$  is sufficiently large. As a differentiable operator  $\mathbb{F}$  is monotone if and only if  $D\mathbb{F} + D\mathbb{F}^\top \succeq 0$  everywhere [114], this shows that  $\mathbb{F}$  is not monotone.

Note that while we provided a two-player example for simplicity, one can easily use the essentially same ideas to construct a non-cocoercive  $n$ -player game satisfying our assumptions with any  $n > 2$ . For example, we can choose  $f_i(x^i; x^{-i}) = \frac{(x^i)^2}{2}\varphi(x^{i+1})$  where we identify  $x^{n+1} = x^1$ .