Decoupled Stochastic Gradient Descent for N-Player Games

Ali Zindari[*]
CISPA[†]
ali.zindari@cispa.de

Parham Yazdkhasti[*]
CISPA[†]
parham.yazdkhasti@cispa.de

Tatjana Chavdarova
CISPA[†]
tatjana.chavdarova@berkeley.edu

Sebastian U. Stich
CISPA[†]
stich@cispa.de

Abstract

We study games with N players where each player aims to minimize their own loss. These games are gaining popularity due to their wide range of applications in machine learning. For instance, minimax optimization problems are a special case of N-player games. Stochastic Gradient Descent (SGD) is among the main methods for solving such games. However, in many distributed game optimization applications, this approach can result in high communication overhead, as each player needs access to the other players' strategies at every time step to compute a gradient. In this paper, we introduce a new optimization paradigm called *Decoupled SGD*. This framework allows individual players to carry out SGD updates independently, with occasional strategy exchanges at predetermined intervals. We analyze the convergence properties of this approach in various scenarios. Primarily, we consider the popular minimax bi-linear game and establish the convergence rate of our method in this setting. We also derive explicit formulas for the optimal length of synchronization intervals and step size. We then provide a general algorithm for N-player games for cases where strategy synchronization is costly. We derive its convergence rate when the resulting operator is strongly monotone. Finally, for minimax optimization problems, we investigate the combination of our Decoupled SGD with classical distributed paradigms, where players have multiple processors/clients and synchronize their strategies sporadically.

1 Introduction

Reinforcement learning (RL) is a versatile and robust framework designed for making decisions under uncertainty. It has gained considerable traction due to its general applicability and powerful problem-solving capabilities. RL has achieved human-level performance in different games like Go and Atari games Silver et al. [2016], Mnih et al. [2015]. Multi-agent systems are one of the important applications of RL, where multiple agents interact within a shared environment. This interaction can be competitive, cooperative, or a mix of both. Multi-agent RL problems are closely related to game theory, as they can be viewed as specific types of multiplayer games where agents interact within a shared environment. In both fields, agents aim to maximize their own rewards, often considering the strategies and actions of other agents to do so. Game theory provides a theoretical framework for understanding the strategic interactions among agents, which is directly applicable to

[*] Equal contribution.

[†] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

multi-agent RL scenarios. Moreover, Many real-world problems in various fields, such as economics and computer science, can be formulated as N-player differentiable games. The goal is to find a set of strategies that no player has an incentive to change their strategy unilaterally. This set is the so-called Nash Equilibrium (NE). Minimax optimization problems, as a special case of N-player games have gained huge attention because of their wide range of applications. This concept appears in areas such as: game theory Von Neumann and Morgenstern [2007], Generative Adversarial Networks (GANs) Goodfellow et al. [2014], adversarial training and robustness Shafahi et al. [2019], Madry et al. [2017], multi-agent RL Li et al. [2019] and adversarial RL Yu et al. [2019]. The most widely used method for finding NE in N-player games is the use of first-order methods such as Gradient Descent Ascent (GDA). A critical assumption in these methods is that each player has access to the strategies of other players, allowing them to compute their gradients accurately and update their strategies accordingly. However, this assumption is not aligned with the realities of multi-agent adversarial RL problems, where agents often do not have direct access to the strategies or parameters of their opponents.

To address this gap, we propose a novel formulation of distributed minimax optimization tailored for N-player games. Our algorithm, proposed in Algorithm 1, effectively computes the NE without violating the realistic constraints discussed. We provide a convergence proof, demonstrating that our algorithm converges to the NE at an exponential rate. This new approach not only distributes the computational load efficiently but also incorporates the practical dynamics of real-world competition.

It might also be the case that each player is on a separate device and these devices are performing online learning. For instance, players can be multiple robots that are cooperating towards accomplishing a goal, each having only access to their own observations of the environment. In this setting, communicating all of the robot's strategies at every time step is very costly and not feasible in real-world problems. All these reasons motivated us to design a new algorithm that is efficient in terms of the number of communications and allows players to avoid sharing their strategies at each step.

1.1 Contributions

- We propose a new algorithm called **Decoupled SGD**, which aims to solve the general N-player games assuming each player has a strongly convex payoff function. With this algorithm, players improve their strategies for K iterations (local steps) without being informed about the most recent strategies of their opponents. This way players only share their parameters every K step which is efficient.

- We first show the convergence of our method for a special simplified class of bi-linear zero-sum games with two players. This case provides more intuition as we have propose an explicit rate based on the parameters of our setting. We show the trade-off between step size and number of local steps and discuss the optimal combination of these two.

- We generalize our convergence rate for the case that there exists N players, each having a strongly convex objective.

- Finally, we provide a convergence rate for our method in the distributed minimax games in which there are only two players having their data distributed across computing nodes.

Note that throughout the paper, when the problem is minimax, we use the name Decoupled (S)GDA for our method and when we are in N-player setting, we use the name Decoupled SGD. The rest of the paper is organized as follows: in Section 2, we discuss some of the related works. In Section 3, we introduce the our notation, In Section 4, we introduce our new formulation and Decoupled SGD algorithm for N-player games. In Section 5, we will provide convergence rates for Decoupled SGD. In Section 6, we extend our method for the distributed minimax games and provide a rate for this setting. Finally, in the last section we provide some experiments to show the effectiveness of our method.

2 Problem formulation

N-player games are defined in the following way:

$$\begin{cases} \min_{\mathbf{x}^1} f_1(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N) \\ \min_{\mathbf{x}^2} f_2(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N) \\ \quad\vdots \\ \min_{\mathbf{x}^N} f_N(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N) \end{cases} \tag{1}$$

Where $\mathcal{X}_n \subseteq \mathbb{R}^{d_n}, \forall n \in [N]$ is a convex set which the function $f_n : \mathcal{X}_1 \times ... \times \mathcal{X}_N \to \mathbb{R}$ is defined on. We recover the well-known zero-sum minimax games as a special case of (1) when $N = 2$ and $f = f_1 = -f_2$. In general, minimax optimization refers to finding the saddle point of objective $f(\mathbf{x}, \mathbf{y})$ where we minimize over $\mathbf{x}$ and maximize over $\mathbf{y}$. This problem can be formulated as:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}). \tag{2}$$

Where $\mathcal{X} \subseteq \mathbb{R}^m$ and $\mathcal{Y} \subseteq \mathbb{R}^n$ are two convex sets on which our function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is defined. Many methods have been proposed for solving the above problems Nouiehed et al. [2019], Korpelevich [1976], Popov [1980], Chavdarova et al. [2020], but the most common way of solving games is to take gradient steps for each player. In zero-sum minimax games, we update the variable $\mathbf{x}$ in the opposite direction of the gradient while updating the variable $\mathbf{y}$ in the direction of the gradient. This method is called Gradient Descent Ascent (GDA). In many real-world scenarios, players in the game can be distributed across several processing units. This leads us to utilize distributed optimization methods in the context of games. However, the most concerning issue in distributed optimization is communication efficiency, as communication is expensive, especially in networks with limited bandwidth. Existing works have studied the case of minimax zero-sum games with two players in the context of distributed learning. However, they often assume that all processors have access to both variables $\mathbf{x}$ and $\mathbf{y}$ and can update them simultaneously. This problem can be formulated as follows:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \frac{1}{M} \sum_{m=1}^{M} f_m(\mathbf{x}, \mathbf{y}). \tag{3}$$

Some works consider the heterogeneous version of the above problem, where f_ms are different, while others assume that all processors have the same f.

3 Related works

We discuss the works on minimax optimization in three categories centralized, decentralized, and distributed. Also we point out to some relevant works on bandits.

Centralized minimax optimization. Many works proposed algorithms for solving minimax optimization on a single machine. Nemirovski [2004], Nesterov [2007] studied the case of convex-concave minimax and proposed a method for this problem that can achieve a rate of $\mathcal{O}(\frac{1}{T})$. Thekumparampil et al. [2019] combined the idea from Nestrov's Accelerated Gradient and mirror-prox and achieved a rate of $\tilde{\mathcal{O}}(\frac{1}{T^2})$ for strongly-convex-strongly-concave functions. Wang and Li [2020] designed an efficient algorithm for general strongly-convex-strongly-concave functions by using the idea from an accelerated proximal point algorithm and can achieve a linear rate. Kovalev and Gasnikov [2022] was the first to propose the optimal method for this class of functions achieving a rate of $\mathcal{O}(\sqrt{\kappa_x\kappa_y} \log \frac{1}{\epsilon})$ which matches the lower bounds in Zhang et al. [2022b], Ibrahim et al. [2020]. Another line of research on minimax optimization [Lee et al., 2024, Zhang et al., 2022a] discovers the effect of alternating in the convergence of GDA.

Decentralized minimax optimization. Decentralized optimization is widely studied for the case of minimization [Xiao and Boyd, 2004, Tsitsiklis, 1984] with the goal of not relying on a central node or server. This idea is also applied to the case of minimax optimization problems. The paper

Liu et al. [2020] is the first who studied non-convex-non-concave decentralized minimax. They also used the idea of optimistic gradient descent and achieved a rate of $\mathcal{O}(\epsilon^{-12})$. In Xian et al. [2021], authors proposed an algorithm called DM-HSGD for non-convex decentralized minimax by utilizing variance reduction and achieved a rate of $\mathcal{O}(\kappa^3\epsilon^{-3})$. Recently, authors in Liu et al. [2023] proposed an algorithm named Precision for the non-convex-strongly-concave objectives which has a two-stage local updates and gives a rate of $\mathcal{O}(\frac{1}{T})$.

Distributed minimax optimization. There is a long track of work for distributed minimization so called Federated Learning (FL) starting with McMahan et al. [2017]. Several works studied FL under different assumptions and data distributions Stich [2018], Koloskova et al. [2020], Karimireddy et al. [2020], Woodworth et al. [2020a,b]. In the context of minimax optimization, there are a few works who studied distributed version of it. The works Deng and Mahdavi [2021], Sharma et al. [2022] proposed rate for different classes of functions in the both heterogeneous and homogeneous regimes. These papers used the formulation 3 which is discussed before. The main difference between these works and ours, is this formulation.

Multiplayer multi-armed bandit. In this class of problems, we have an environment with N players trying to solve a Multi-Armed Bandit (MAB) problems while collaborating with other players. The goal is to maximize the cumulative reward or to minimize the regret. The work Wang et al. [2020] proposed a new method for solving distributed MAB problems that can achieve the same regret bound as in the centralized setting. Another work Agarwal et al. [2022] considered a regime in which the goal is not only to minimize the regret, but also the number of communication and the number of bits used in each communication. Another variation of distributed MAP is to collaboratively identify the arm with the highest average reward. Authors in Mitra et al. [2021] proposed Fed-SEL which is a communication efficient method that benefits from high heterogeneity of arms. Chen et. al Chen et al. [2023] studied MAB assuming players have different speeds in decision making and proposed a new protocol to tackle this issue.

4 Setting and preliminaries

In this section, we start by introducing the notations and definitions that will be used frequently throughout this paper.

4.1 Notations and basic definitions

N-player games. We use $\mathbf{x}_k^{n,r}$ as the parameters of player n at some round r and after k local steps. Also we use $\mathbf{x}_k^{-n,r}$ to denote the concatenation of all players' parameters excluding player n. The concatenation of all players' parameters is shown by $\mathbf{z}_k^r$. Moreover, we define the operator $G(\mathbf{z})$ as follows:

$$G(\mathbf{z}) := \begin{pmatrix} \nabla_{\mathbf{x}^1} f_1(\mathbf{x}^1, \ldots, \mathbf{x}^N) \\ \vdots \\ \nabla_{\mathbf{x}^N} f_N(\mathbf{x}^1, \ldots, \mathbf{x}^N) \end{pmatrix} \tag{4}$$

The NE in N-player games Bravo et al. [2018] is defined as:

$$f_n(\mathbf{x}_\star^n; \mathbf{x}_\star^{-n}) \leqslant f_n(\mathbf{x}^n; \mathbf{x}_\star^{-n}), \quad \forall \mathbf{x}^n \in \mathcal{X}_n, n \in [N] \tag{5}$$

Where $\mathbf{x}_\star^n$ is the strategy of player n at NE. The point $\mathbf{z}^\star := \left[\mathbf{x}_\star^1; \ldots; \mathbf{x}_\star^N\right]$ has to satisfy the following condition in an unconstrained game:

$$G(\mathbf{z}^\star) = 0 \tag{6}$$

For the special case of $N = 2$ and $f = f_1 = -f_2$, we recover the definition of well-known minimax games. In these games, we aim to find the saddle point $(\mathbf{x}^\star, \mathbf{y}^\star)$ which has to satisfy the following property:

$$f(\mathbf{x}^\star, \mathbf{y}) \leqslant f(\mathbf{x}^\star, \mathbf{y}^\star) \leqslant f(\mathbf{x}, \mathbf{y}^\star), \forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y}. \tag{7}$$

5 Decoupled SGD for N-player games

In this section, we introduce our Decoupled SGD algorithm for N-player games. Our algorithm has a round-wise structure meaning that at the beginning of some round r, player n sends his parameters to all other players and also receives the parameters of all other players. Then all players start taking SGD updates for K steps, only updating their own strategies. The advantage of our method is that each player **doesn't need to compute the gradient** of his payoff function with respect to the strategies of other players. It only requires an outdated version of their parameters which has been received at the beginning of the round, leading to communication efficiency. In line 8 of our method in Algorithm 1, communication can be done peer-to-peer or through a central server.

Algorithm 1 Decoupled SGD for N-player games

```
 1: Input: step size $\gamma$, initialization $\mathbf{x}_0^1, \ldots, \mathbf{x}_0^N$
 2: for $r \in \{1, \ldots, R\}$ do
 3:     for $k \in \{0, \ldots, K-1\}$ do
 4:         for $n \in \{1, \ldots, N\}$ in parallel do
 5:             Update local model $\mathbf{x}_{k+1}^{n,r} \leftarrow \mathbf{x}_k^{n,r} - \gamma \nabla f_n(\mathbf{x}_k^{n,r}; \mathbf{x}_0^{-n,r})$
 6:         end for
 7:     end for
 8:     Communicate $\left[\mathbf{x}_K^{1,r}, \ldots, \mathbf{x}_K^{N,r}\right]^\top$ to all players
 9: end for
10: Output: $\mathbf{x}_K^{1,R}, \ldots, \mathbf{x}_K^{N,R}$
```

In the following, we introduce the assumptions that are crucial for the convergence of our Algorithm.

Assumption 1. *The variance of stochastic gradient on function f_n is uniformly upper bounded.*

$$\mathbb{E}_{\xi_n}\left[\left\|\nabla_{\mathbf{x}^n} f_n(\mathbf{x}^1, \ldots, \mathbf{x}^N; \xi_n) - \nabla_{\mathbf{x}^n} f_n(\mathbf{x}^1, \ldots, \mathbf{x}^N)\right\|^2\right] \leqslant \sigma^2. \tag{8}$$

Assumption 2. *An operator $G : \mathbb{R}^d \to \mathbb{R}^n$ is called to be L-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, there exist a constant $L > 0$ such that:*

$$\|G(\mathbf{x}) - G(\mathbf{y})\| \leqslant L \|\mathbf{x} - \mathbf{y}\|. \tag{9}$$

Assumption 3. *An operator $G : \mathbb{R}^d \to \mathbb{R}^d$ is called to be μ strongly monotone if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, there exist a constant $\mu > 0$ such that:*

$$\left\langle G(\mathbf{x}) - G(\mathbf{y}), \mathbf{x} - \mathbf{y} \right\rangle \geqslant \mu \|\mathbf{x} - \mathbf{y}\|^2. \tag{10}$$

6 Convergence analysis

In this section, we provide convergence guarantees for our proposed methods. We start this section by providing a rate for the case of bi-linear zero-sum minimax games. This gives more intuition about our method as we have a closed form for the class of bi-linear games. Next, we generalize our results to the class of N-player games with strongly convex objectives.

6.1 Decoupled GDA (Alg. 1) for bi-linear games

A simplified bi-linear game can be defined as:

$$f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^\top(\omega \mathbf{I})\mathbf{x} - \frac{1}{2}\mathbf{y}^\top(\omega \mathbf{I})\mathbf{y} + \mathbf{x}^\top \mathbf{C}\mathbf{y} \tag{11}$$

Where $\omega \in \mathbb{R}$ and $\mathbf{C} \in \mathbb{R}^{d \times d}$ is a symmetric matrix defining the interactive part of the game.

Theorem 1. *For any K, R after running Decoupled GDA for a total of $T = KR$ iterations on the problems in the form of (11), with a learning rate of $\gamma \leqslant \frac{1}{\omega}$, we have a last iterate convergence rate of:*

$$\left\|\mathbf{z}_K^R - \mathbf{z}^\star\right\|^2 \leqslant \left((1-\gamma\omega)^{2K} + \left((1-\gamma\omega)^K - 1\right)^2 \omega^{-2} \lambda_{\max}^2(\mathbf{C})\right)^R \left\|\mathbf{z}_0^0 - \mathbf{z}^\star\right\|^2 \tag{12}$$

Where $\lambda_{\max}(\mathbf{C})$ refers to the maximum eigenvalue of the matrix $\mathbf{C}$.

Discussion To ensure the convergence of our method, the coefficient of $\left\|\mathbf{z}_0^0 - \mathbf{z}^\star\right\|^2$ should be less than one. This gives us some conditions on the steps size and number of local steps. In general, the more number of local steps we take, the smaller step size should be used.

Corollary 2. *There exists an optimal combination of number of local steps and step size $(\gamma^\star, K^\star)$ for Theorem 1, which gives the optimal rate of:*

$$\left\|\mathbf{z}_K^R - \mathbf{z}^\star\right\|^2 \leqslant \left(\frac{\lambda_{\max}^2(\mathbf{C})}{\omega^2 + \lambda_{\max}^2(\mathbf{C})}\right)^R \left\|\mathbf{z}_0^0 - \mathbf{z}^\star\right\|^2 \tag{13}$$

From (13) we can see that the when $\lambda_{\max}^2(\mathbf{C}) \to 0$, which means that the game has no interactive part, we converge in one step which is expected. On the other hand, if $\omega \to 0$, we never converge as we have a game with only the term $\mathbf{x}^\top \mathbf{C}\mathbf{y}$. It's widely known that GDA doesn't have a last iterate convergence guarantee for this type of game.

6.2 N-Player games

Theorem 3. *For any $K, R, L > 0, \mu > 0$ after running Decoupled SGD for a total of $T = KR$ iterations on the problems in the form of (1) with a learning rate of $\gamma \leqslant \frac{\mu}{32L^2KN}$, assuming that $\left\|\mathbf{z}_0^0 - \mathbf{z}^\star\right\|^2 \leqslant B^2$ and operator G is strongly monotone, we have the following convergence rate up to some logarithmic factors:*

$$\mathbb{E}\left\|\mathbf{z}_K^R - \mathbf{z}^\star\right\|^2 = \tilde{\mathcal{O}}\left(B^2 \exp\left(-\frac{\mu^2}{NL^2}R\right) + \frac{\sigma^2}{\mu^2 KR}\right)$$

Discussion From the Theorem 3 it's clear that $R = \omega(N)$ for the convergence. This comes from the fact that we need to scale the step size with the number of players. As this number increases, we have more outdated parameters being used in our update rule which results in a higher error so it's intuitive to choose a smaller learning rate when the number of players is very large.

7 Decoupled SGDA for distributed zero-sum minimax games

In this section, we study an extension of our algorithm for distributed setting. For simplicity and in order to be aligned with other works Deng and Mahdavi [2021], Sharma et al. [2022], we consider two-player zero-sum minimax games. Our results for the distributed setting can be extended to the N-player case.

Notations In this setting, we assume that each player's data is distributed across M processors. So each processor has access to a function $f_m(\mathbf{x}, \mathbf{y})$ on which it can perform gradient steps. The distribution of data across processors can be either homogeneous or heterogeneous. In the heterogeneous regime, which is the case of study in this paper, each processor holds a different payoff function. To measure this difference, we use the following assumption:

Assumption 4. *There exists a constant $\zeta_\star$ satisfying the following inequality in distributed minimax games:*

$$\max\left\{\sup_m \left\|\nabla_{\mathbf{x}} f_m(\mathbf{z}^\star)\right\|^2, \sup_m \left\|\nabla_{\mathbf{y}} f_m(\mathbf{z}^\star)\right\|^2\right\} \leqslant \zeta_\star^2 \tag{14}$$

Where $\mathbf{z}^\star := [\mathbf{x}^\star; \mathbf{y}^\star]$ is the saddle point.

We denote $\mathbf{x}_k^{m,r}$ and $\mathbf{y}_k^{m,r}$ as the parameters of players $\mathbf{x}$ and $\mathbf{y}$ on client m in some round r after k local steps. The concatenation of $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\mathbf{z}$. We also operators $G(\mathbf{z}), G_m(\mathbf{z})$ are defined as follows:

$$G_m(\mathbf{z}) := \begin{pmatrix} \nabla_{\mathbf{x}} f_m(\mathbf{z}) \\ -\nabla_{\mathbf{y}} f_m(\mathbf{z}) \end{pmatrix}, \quad G(\mathbf{z}) := \begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{z}) \\ -\nabla_{\mathbf{y}} f(\mathbf{z}) \end{pmatrix} \tag{15}$$

Note that in general $G_m(\mathbf{z}^\star) \neq 0$. In Algorithm 2, we discuss the distributed version of our method, where two players $\mathbf{x}$ and $\mathbf{y}$ have their data distributed across M processors each. At every round, each set of processors update their local models while having access to an outdated version of the other opponent parameters which was received at the beginning of the round. By the end of the round, both set of $\mathbf{x}$ and $\mathbf{y}$ processors send the their parameters to a central server which will compute the average of the parameters and send them back to all processors.

Algorithm 2 Decoupled SGDA for 2-player distributed minimax games

```
**Input:** step size $\gamma$, initialization $\mathbf{x}_0, \mathbf{y}_0$
**Initialize:** $\forall m \in [M], \mathbf{x}_0^{r,0} \leftarrow \mathbf{x}_0, \quad \mathbf{y}_0^{r,0} \leftarrow \mathbf{y}_0$
**for** $r \in \{1, \ldots, R\}$ **do**
    $\forall m \in [M], \mathbf{x}_0^{m,r} \leftarrow \bar{\mathbf{x}}_0^r, \quad \mathbf{y}_0^{m,r} \leftarrow \bar{\mathbf{y}}_0^r$
    **for** $k \in \{0, \ldots, K-1\}$ **do**
        **for** $m \in \{1, \ldots, M\}$ **in parallel do**
            Update local model $\mathbf{x}_{k+1}^{m,r} \leftarrow \mathbf{x}_k^{m,r} - \gamma \nabla f(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r})$
            Update local model $\mathbf{y}_{k+1}^{m,r} \leftarrow \mathbf{y}_k^{m,r} + \gamma \nabla f(\mathbf{x}_0^{m,r}, \mathbf{y}_k^{m,r})$
        **end for**
    **end for**
    $\bar{\mathbf{x}}_0^{r+1} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}_K^{m,r}, \quad \bar{\mathbf{y}}_0^{r+1} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \mathbf{y}_K^{m,r}$
    **Communicate** $\bar{\mathbf{x}}_K^r$ to all processors with $\mathbf{y}$ player and $\bar{\mathbf{y}}_K^r$ to all processors with $\mathbf{x}$ player
**end for**
**Output:** $\bar{\mathbf{x}}_K^R, \bar{\mathbf{y}}_K^R$
```

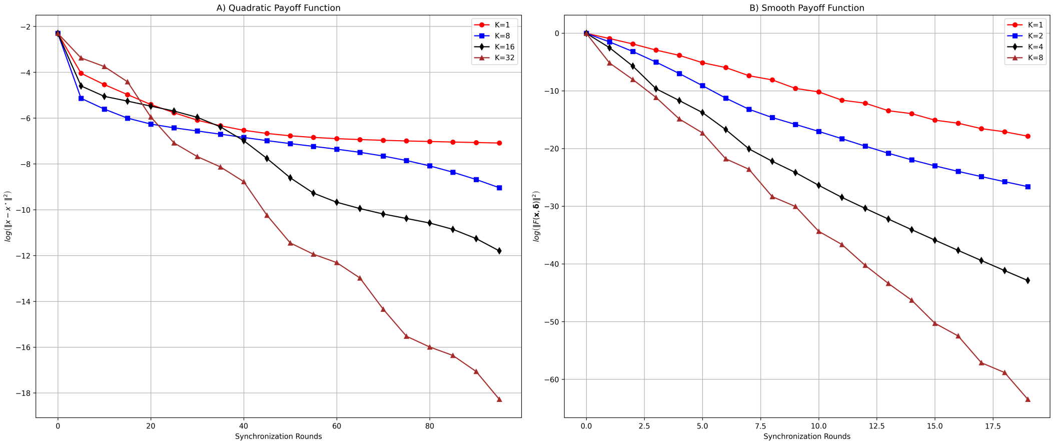

Figure 1: Convergence rates for different synchronization rounds in two scenarios: (A) Quadratic Payoff Function and (B) Smooth Payoff Function. In both cases, increasing the synchronization rounds (K) results in faster convergence, with higher K values showing a more rapid decline in error.

Theorem 4. *For any $K, R, L > 0, \mu > 0$ after running Decoupled SGDA for a total of $T = KR$ iterations on the problems in the form of (2) in a distributed setting with a learning rate of $\gamma \leqslant \frac{\mu}{32L^2K}$, assuming that $\|\mathbf{z}_0 - \mathbf{z}^\star\|^2 \leqslant B^2$ and operator G is strongly monotone, we have the following convergence rate:*

$$\mathbb{E}\left\|\bar{\mathbf{z}}_K^R - \mathbf{z}^\star\right\|^2 = \tilde{\mathcal{O}}\left(B^2 \exp\left(-\frac{\mu^2}{L^2}R\right) + \frac{L^2\zeta_\star^2}{\mu^4R^2} + \frac{L^2\sigma^2}{\mu^4KR^2} + \frac{\sigma^2}{\mu^2MKR}\right)$$

Discussion The first term in our rate benefits from an exponential decrease. The second term is affected by $\zeta_\star$ which comes from the fact that we assumed data is not identically distributed across the processors. We can recover the result for identically distributed data by just setting $\zeta_\star = 0$.

8 Experiments

In this section, we show some of the theoretical properties of our proposed method with experiments[3]. We used a MacBook m2 laptop for running the experiments.

[3] https://anonymous.4open.science/r/Decoupled-Stochastic-Gradient-Descent-for-N--Player-Games-C3C5/

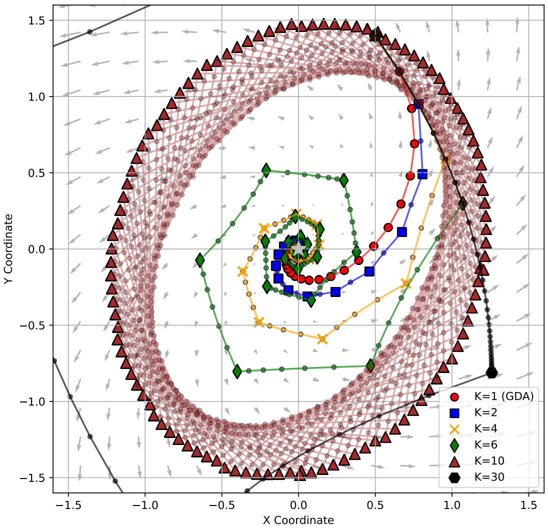

Figure 2: This figure illustrates players' strategies trajectories when different local steps are used. We can see that Decoupled GDA diverges for K larger than some threshold value. This value for this problem set is $K = 10$.

8.1 Finding the saddle point of bi-linear games

In the first experiment, we demonstrate the validity of our claim that local steps can accelerate convergence in terms of communication rounds. As shown on the left side of Figure 1, we consider a simple bi-linear game formulated as

$$f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \frac{1}{2}\mathbf{y}^\top \mathbf{B}\mathbf{y} + \mathbf{x}^\top \mathbf{C}\mathbf{y}$$

to investigate the impact of the number of local steps on the convergence rate. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are positive semi-definite matrices, randomly generated in $\mathbb{R}^{5\times5}$. Each of these random matrices is then normalized to have a fixed maximum eigenvalue. From the figure, it is evident that incorporating local steps results in a significant speed-up in convergence.

8.2 Communication efficiency of Decoupled SGD for functions beyond bi-linear

Here we present a synthetic minimax optimization problem that doesn't have a bi-linear form then we show that our method is communication efficient on this function. The function we consider is defined as follows:

$$\min_{\mathbf{x}} \max_{\delta} \frac{1}{N} \left\|\sigma(\mathbf{A}(\mathbf{x}+\delta)) - \mathbf{y}\right\|^2 + \lambda_1 \left\|\mathbf{x}\right\|^2 - \lambda_2 \left\|\delta\right\|^2$$

Where $\mathbf{x}, \mathbf{y}, \delta \in \mathbb{R}^5$, $\mathbf{A} \in \mathbb{R}^{100\times5}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ are just regularization parameters. In the above equation, σ is the sigmoid function. It prevents the above equation from having a bi-linear form. This type of problem formulation can be found in the context of adversarial training.The right-hand side of Figure 1 illustrates the norm of the operator after running GD and Decoupled GD with their optimal step size. The algorithm is run for different step sizes for each value of K, and the best-performing one is chosen to be illustrated here. It is clear that for a fixed accuracy, our algorithm with $K = 50$ had the best performance and required the least number of communication rounds.

9 Conclusion

In this work, we introduced a novel approach for solving N-player games, which we believe is more practical and suitable for real-world applications. The primary advantage of our method lies in its ability to operate without requiring access to all players' strategies at each step. This feature provides greater autonomy to players who may prefer not to continuously share their parameters. Furthermore, we have theoretically demonstrated that our method achieves convergence even when utilizing outdated gradients. There might be some room for improvement by utilizing newer approaches in minimax optimization such as Extra Gradient method which we leave for future works.

References

M. Agarwal, V. Aggarwal, and K. Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *Journal of Machine Learning Research*, 23(212):1–24, 2022.

M. Bravo, D. Leslie, and P. Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

T. Chavdarova, M. Pagliardini, S. U. Stich, F. Fleuret, and M. Jaggi. Taming gans with lookahead-minmax. *arXiv preprint arXiv:2006.14567*, 2020.

Y.-Z. J. Chen, L. Yang, X. Wang, X. Liu, M. Hajiesmaili, J. C. Lui, and D. Towsley. On-demand communication for asynchronous multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3903–3930. PMLR, 2023.

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

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

A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International conference on machine learning*, pages 4583–4593. PMLR, 2020.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

J. Lee, H. Cho, and C. Yun. Fundamental benefit of alternating updates in minimax optimization. *arXiv preprint arXiv:2402.10475*, 2024.

S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4213–4220, 2019.

M. Liu, W. Zhang, Y. Mroueh, X. Cui, J. Ross, T. Yang, and P. Das. A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33: 11056–11070, 2020.

Z. Liu, X. Zhang, S. Lu, and J. Liu. Precision: Decentralized constrained min-max learning with low communication and sample complexities. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 191–200, 2023.

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

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

A. Mitra, H. Hassani, and G. Pappas. Exploiting heterogeneity in robust federated best-arm identification. *arXiv preprint arXiv:2109.05700*, 2021.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

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

L. D. Popov. A modification of the arrow-hurwitz method of search for saddle points. *Mat. Zametki*, 28(5):777–784, 1980.

A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

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

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

S. U. Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

S. U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

J. N. Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

R. S. Varga. Springer series in computational mathematics. 2004.

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.

P.-A. Wang, A. Proutiere, K. Ariu, Y. Jedra, and A. Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

Y. Wang and J. Li. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.

B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020a.

B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020b.

W. Xian, F. Huang, Y. Zhang, and H. Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34:25865–25877, 2021.

L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53 (1):65–78, 2004.

L. Yu, J. Song, and S. Ermon. Multi-agent adversarial inverse reinforcement learning. In *International Conference on Machine Learning*, pages 7194–7201. PMLR, 2019.

G. Zhang, Y. Wang, L. Lessard, and R. B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 7659–7679. PMLR, 2022a.

J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1):901–935, 2022b.

A Appendix

Lemma 5. *For a convex function f we have:*

$$f\left(\frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_m\right) \leqslant \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x}_m)\,. \tag{16}$$

Lemma 6. *For a set of M vectors $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_M \in \mathbb{R}^d$ we have:*

$$\left\|\sum_{m=1}^{M}\mathbf{a}_m\right\| \leqslant \sum_{m=1}^{M}\|\mathbf{a}_m\|\ . \tag{17}$$

Lemma 7. *For a set of M vectors $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_M \in \mathbb{R}^d$ we have:*

$$\left\|\sum_{m=1}^{M}\mathbf{a}_m\right\|^2 \leqslant M\sum_{m=1}^{M}\|\mathbf{a}_m\|^2\ . \tag{18}$$

Lemma 8. *For two arbitrary vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $\forall\gamma > 0$ we have:*

$$\|\mathbf{a}+\mathbf{b}\|^2 \leqslant (1+\gamma)\|\mathbf{a}\|^2 + (1+\gamma^{-1})\|\mathbf{b}\|^2\ . \tag{19}$$

Lemma 9. *Let Assumption 1 holds. Then we have:*

$$\mathbb{E}_{\xi_m}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla_{\mathbf{x}} f_m(\mathbf{x},\mathbf{y},\xi_m) - \frac{1}{M}\sum_{m=1}^{M}\nabla_{\mathbf{x}} f_m(\mathbf{x},\mathbf{y})\right\|^2 \leqslant \frac{\sigma^2}{M}\,. \tag{20}$$

The same argument holds for gradient with respect to $\mathbf{y}$.

Lemma 10 (Duality gap). *For a strongly-convex-strongly-concave function $f(\mathbf{x}, \mathbf{y})$ we have:*

$$-\langle F(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star\rangle \leqslant -\Big[f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t)\Big] - \frac{\mu}{2}\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \tag{21}$$

Proof. We know that our objective is strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$ which implies:

$$
\begin{aligned}
-\langle \nabla_x f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle &\leqslant f(\mathbf{x}^\star, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t) - \frac{\mu}{2} \left\| \mathbf{x}_t - \mathbf{x}^\star \right\|^2 \\
-\langle \nabla_y f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}^\star - \mathbf{y}_t \rangle &\leqslant f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}^\star) - \frac{\mu}{2} \left\| \mathbf{y}_t - \mathbf{y}^\star \right\|^2
\end{aligned}
$$

Summing up the above inequalities gives us:

$$
-\langle \nabla_x f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle - \langle \nabla_y f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}^\star - \mathbf{y}_t \rangle \leqslant f(\mathbf{x}^\star, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}^\star) - \frac{\mu}{2} \left\| \mathbf{x}_t - \mathbf{x}^\star \right\|^2 - \frac{\mu}{2} \left\| \mathbf{y}_t - \mathbf{y}^\star \right\|^2
$$

By re-writing the above expression based on $\mathbf{z}$ we have:

$$
-\langle F(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle \leqslant -\Big[f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \Big] - \frac{\mu}{2} \left\| \mathbf{z}_t - \mathbf{z}^\star \right\|^2
$$

□

Lemma 11. *Let $\{r_t\}_{t \geqslant 0}$ be a non-negative sequence of numbers that satisfy*

$$
r_{t+1} \leqslant (1 - a\gamma) r_t + \frac{b}{K} \gamma \sum_{i=\max\{0, t-K+1\}}^{t} r_i + c\gamma^2 \,,
$$

for constants $a > 0$, $b, c \geqslant 0$ and integer $K \geqslant 1$ and a parameter $\gamma \geqslant 0$, such that $a\gamma \leqslant \frac{1}{K}$. If $b \leqslant \frac{a}{4}$, then it holds

$$
r_t \leqslant \Big(1 - \frac{a}{2}\gamma\Big)^t r_0 + \frac{2c}{a}\gamma \,. \tag{22}
$$

Proof. By assumption on r_t:

$$
r_{t+1} \leqslant \Big(1 - \frac{a\gamma}{2}\Big) r_t - \frac{a\gamma}{2} r_t + \frac{b}{K}\gamma \sum_{i=\max\{0, t-K+1\}}^{t} r_i + c\gamma^2 \,,
$$

and by unrolling the recursion:

$$
\begin{aligned}
r_{t+1} &\leqslant \Big(1 - a\frac{\gamma}{2}\Big)^t r_0 + \sum_{i=0}^{t} \Big(1 - \frac{a\gamma}{2}\Big)^{t-i} \left[-\frac{a\gamma}{2} r_i + \frac{b}{K}\gamma \sum_{j=\max\{0, i-K+1\}}^{i} r_j \right] + \sum_{i=0}^{t} \Big(1 - \frac{a\gamma}{2}\Big)^{t-i} c\gamma^2 \\
&\leqslant \Big(1 - \frac{a\gamma}{2}\Big)^t r_0 + \sum_{i=0}^{t} \Big(1 - \frac{a\gamma}{2}\Big)^{t-i} \left[-\frac{a\gamma}{2} r_i + \frac{b}{K}\gamma \sum_{j=\max\{0, i-K+1\}}^{i} r_j \right] + \frac{2c}{a}\gamma \\
&= \Big(1 - a\frac{\gamma}{2}\Big)^t r_0 + \sum_{i=0}^{t} \Big(1 - \frac{a\gamma}{2}\Big)^{t-i} \left[-\frac{a\gamma}{2} r_i + \frac{b}{K}\gamma \sum_{j=\max\{0, i-K-1\}}^{i} \Big(1 - \frac{a\gamma}{2}\Big)^{i-j} r_i \right] + \frac{2c}{a}\gamma
\end{aligned}
$$

where we used $\sum_{i=0}^{t} (1 - \frac{a\gamma}{2})^i \leqslant \frac{2}{a\gamma}$ (for $(\frac{a\gamma}{2}) < 1$) for the second inequality.

By estimating

$$
\begin{aligned}
-\frac{a\gamma}{2} r_i + \frac{b}{K}\gamma \sum_{j=\max\{0, i-K-1\}}^{i} (1 - \frac{a\gamma}{2})^{i-j} r_i &\leqslant -\frac{a\gamma}{2} r_i + \frac{b}{K}\gamma \sum_{j=\max\{0, i-K-1\}}^{i} \Big(1 - \frac{a\gamma}{2}\Big)^{1-K} r_i \\
&\leqslant -\frac{a\gamma}{2} r_i + b\gamma r_i \Big(1 - \frac{a\gamma}{2}\Big)^{1-K} r_i \\
&\leqslant -\frac{a\gamma}{2} r_i + 2b\gamma r_i \leqslant 0 \,,
\end{aligned}
$$

with and $(1 - \frac{a\gamma}{2})^{1-K} \leqslant 2$ for $a\gamma \leqslant \frac{1}{K}$, and the assumption $b \leqslant \frac{a}{4}$ (and $r_i \geqslant 0$).

The validity of the inequality, $(1-\frac{a\gamma}{2})^{1-K} \leqslant 2$ for $a\gamma \leqslant \frac{1}{K}$ can be shown in the following way:

$$\left(1-\frac{a\gamma}{2}\right)^{1-K} \leqslant \left(1-\frac{a\gamma}{2}\right)^{-K} \leqslant e^{\frac{a\gamma K}{2}}$$

For the last inequality above we used the approximation $(1-x)^{-n} \leqslant e^{nx}$ for $x \geqslant 0$ and $n \geqslant 0$: Given that $a\gamma \leqslant \frac{1}{K}$, we have:

$$e^{\frac{a\gamma K}{2}} \leqslant e^{\frac{1}{2}}.$$

Thus, we have

$$\left(1-\frac{a\gamma}{2}\right)^{1-K} \leqslant 2$$

Going back to the main proof, we conclude

$$r_{t+1} \leqslant \left(1-\frac{a\gamma}{2}\right)^{t} r_0 + \frac{2c}{a}\gamma\,.$$

as claimed. □

Lemma 12 (Gershgorin's Theorem for Block Matrices Varga [2004])**.** *Consider $\mathcal{A} = (A_{ij}) \in \mathbb{R}^{dn\times dn}$ where $A_{ij} \in \mathbb{R}^{d\times d}$. Suppose $\sigma(\cdot)$ is the spectrum of a matrix. If we denote*

$$G_i \triangleq \sigma\left(A_{ii}\right) \cup \left\{\lambda \notin \sigma\left(A_{ii}\right) : \left\|\left(A_{ii}-\lambda I_d\right)^{-1}\right\|^{-1} \leqslant \sum_{j=1, j\neq i}^{n} \|A_{ij}\|\right\}$$

then

$$\sigma(\mathcal{A}) \in \bigcup_{i=1}^{n} G_i$$

Theorem 1 means the eigenvalue of $\mathcal{A}$ either equals $\sigma\left(A_{ii}\right)$ or in that specific region. Corollary 1. If A_{ii} is symmetric, then the region can be specifically expressed as

$$G_i \triangleq \sigma\left(A_{ii}\right) \cup \left\{\bigcup_{k=1}^{n} C\left(\lambda_k\left(A_{ii}\right), \sum_{j=1, j\neq i}^{n} \|A_{ij}\|\right)\right\}$$

Where $C(*,*)$ donotes a disk

$$C(c,r) = \{\lambda : \|\lambda - c\| \leqslant r\}$$

As shown above G_i contains d circles centered at all the eigenvalues of A_{ii}.

A.1 Consensus error

In this work when we use the term consensus error, we mean the error that is either caused by (1): The use of outdated gradients in our algorithm. (2) The deviation of iterates from their average when we have M processors for each player.

A.1.1 Consensus error in N-player games

In this setting, we only have the error caused by the use of outdated gradients on each player. We define this error for this setting as follows:

$$\Phi(\mathbf{x}_k^{n,r}) := \sum_{n=1}^{N} \|\mathbf{x}_k^{n,r} - \mathbf{x}_0^{n,r}\|^2$$

A.1.2 Consensus error in two-player distributed minimax games

In this setting, we have both errors related to the use of outdated gradients and deviation from the average iterates. Total error is the sum of both errors. We define the consensus error in this setting as follows:

$$\begin{aligned}
&\Psi(\mathbf{x}_k^{m,r}) := \frac{1}{M}\sum_{m=1}^{M}\left\|\mathbf{x}_k^{m,r}-\bar{\mathbf{x}}_k^r\right\|^2, \quad \Psi(\mathbf{y}_k^{m,r}) := \frac{1}{M}\sum_{m=1}^{M}\left\|\mathbf{y}_k^{m,r}-\bar{\mathbf{y}}_k^r\right\|^2 \\
&\Phi(\bar{\mathbf{x}}_k^r) := \left\|\bar{\mathbf{x}}_0^r-\bar{\mathbf{x}}_k^r\right\|^2, \quad \Phi(\bar{\mathbf{y}}_k^r) := \left\|\bar{\mathbf{y}}_0^r-\bar{\mathbf{y}}_k^r\right\|^2 \\
&\Psi(\mathbf{z}_k^{m,r}) = \Psi(\mathbf{x}_k^{m,r}) + \Psi(\mathbf{y}_k^{m,r}), \quad \Phi(\bar{\mathbf{z}}_k^r) = \Phi(\bar{\mathbf{x}}_k^r) + \Phi(\bar{\mathbf{y}}_k^r)
\end{aligned}$$

The total consensus error can be computed by summing both errors with respect to $\mathbf{x}$ and $\mathbf{y}$:

$$\text{Consensus error} := \underbrace{\Psi(\mathbf{x}_k) + \Psi(\mathbf{y}_k)}_{\text{error caused by multiple clients}} + \underbrace{\Phi(\mathbf{x}_k) + \Phi(\mathbf{y}_k)}_{\text{error caused by outdated gradients}}$$

In the following, the upper bound for consensus error in different settings will be discussed. Note that in the case of multi client, we get different upper bounds based on the assumption on data heterogeneity.

Lemma 13 (Consensus error for M clients and two players in heterogeneous setting). *After running Decoupled Local SGDA for k local steps at some round r with a step-size of $\gamma \leqslant \frac{\mu}{32L^2K}$, the error $\Psi(\mathbf{z}_k^{m,r}) + \Phi(\bar{\mathbf{z}}_k^r)$ can be upper bounded as follows: After running Decoupled Local SGDA for k local steps at some round r with a step-size of $\gamma \leqslant \frac{\mu}{32L^2K}$, the error $\Psi(\mathbf{z}_k^{m,r}) + \Phi(\bar{\mathbf{z}}_k^r)$ can be upper bounded as follows:*

$$\mathbb{E}[\Psi(\mathbf{z}_k^{m,r}) + \Phi(\bar{\mathbf{z}}_k^r)] \leqslant \sum_{i=1}^{K}\frac{\mu^2}{8KL^2}\left\|\bar{\mathbf{z}}_i^r - \mathbf{z}^\star\right\|^2 + 32K^2\gamma^2\zeta_\star^2 + \frac{2K\gamma^2\sigma^2}{M} + 2K\gamma^2\sigma^2 \tag{23}$$

Proof.

$$
\begin{aligned}
&\mathbb{E}[\Psi(\mathbf{x}_{k+1}^{m,r}) + \Phi(\bar{\mathbf{x}}_{k+1}^{r})] \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \mathbf{x}_k^{m,r} - \gamma \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}; \xi_m) - \bar{\mathbf{x}}_k^{r} + \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}; \xi_m) \right\|^2 + \\
&\qquad\qquad \mathbb{E}\left\| \bar{\mathbf{x}}_0^{r} - \bar{\mathbf{x}}_k^{r} + \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}; \xi_m) \right\|^2 \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \mathbf{x}_k^{m,r} - \gamma \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) - \bar{\mathbf{x}}_k^{r} + \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) \right\|^2 + \\
&\qquad\qquad \mathbb{E}\left\| \bar{\mathbf{x}}_0^{r} - \bar{\mathbf{x}}_k^{r} + \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&\leqslant \left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + \frac{2K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) - \frac{1}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) \right\|^2 + \\
&\qquad\qquad \frac{2K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&\leqslant \left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + \frac{4K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&= \left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + \\
&\qquad\qquad \frac{4K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) - \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) + \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&\leqslant \left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + \frac{8K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) - \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) \right\|^2 + \\
&\qquad\qquad \frac{8K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&\leqslant \left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + 8KL^2\gamma^2\, \mathbb{E}[\Psi(\mathbf{x}_k^{m,r})] + 8KL^2\gamma^2\, \mathbb{E}[\Phi(\bar{\mathbf{y}}_k^{r})] + \\
&\qquad\qquad \frac{8K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&= \left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + 8KL^2\gamma^2\, \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{y}}_k^{r})] + \\
&\qquad\qquad \frac{8K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) - \nabla_{\mathbf{x}} f_m(\mathbf{x}^{\star}, \mathbf{y}^{\star}) + \nabla_{\mathbf{x}} f_m(\mathbf{x}^{\star}, \mathbf{y}^{\star}) \right\|^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2 \\
&\left(1 + \frac{1}{K}\right) \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{x}}_k^{r})] + 8KL^2\gamma^2\, \mathbb{E}[\Psi(\mathbf{x}_k^{m,r}) + \Phi(\bar{\mathbf{y}}_k^{r})] + \\
&\qquad\qquad \frac{16K\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\| \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^{r}, \bar{\mathbf{y}}_k^{r}) - \nabla_{\mathbf{x}} f_m(\mathbf{x}^{\star}, \mathbf{y}^{\star}) \right\|^2 + 16K\gamma^2\zeta_{\star}^2 + \frac{\gamma^2\sigma^2}{M} + \gamma^2\sigma^2
\end{aligned}
$$

we continue:

$$\mathbb{E}[\Psi(\mathbf{x}^{m,r}_{k+1}) + \Phi(\bar{\mathbf{x}}^r_{k+1})] \leqslant \left(1+\frac{1}{K}\right)\mathbb{E}[\Psi(\mathbf{x}^{m,r}_k) + \Phi(\bar{\mathbf{x}}^r_k)] + 8KL^2\gamma^2\,\mathbb{E}[\Psi(\mathbf{x}^{m,r}_k)+\Phi(\bar{\mathbf{y}}^r_k)] + \\ 16KL^2\gamma^2\,\mathbb{E}\left\|\bar{\mathbf{z}}^r_k - \mathbf{z}^\star\right\|^2 + 16K\gamma^2\zeta_\star^2 + \frac{\gamma^2\sigma^2}{M}+\gamma^2\sigma^2$$

After doing the same computation with respect to $\mathbf{y}$ we get:

$$\begin{aligned}&\mathbb{E}[\Psi(\mathbf{y}^{m,r}_{k+1}) + \Phi(\bar{\mathbf{y}}^r_{k+1})]\\ &\leqslant \left(1+\frac{1}{K}\right)\mathbb{E}[\Psi(\mathbf{y}^{m,r}_k) + \Phi(\bar{\mathbf{y}}^r_k)] + 8KL^2\gamma^2\,\mathbb{E}[\Psi(\mathbf{y}^{m,r}_k)+\Phi(\bar{\mathbf{x}}^r_k)] + \\ &\qquad 16KL^2\gamma^2\,\mathbb{E}\left\|\bar{\mathbf{z}}^r_k - \mathbf{z}^\star\right\|^2 + 16K\gamma^2\zeta_\star^2 + \frac{\gamma^2\sigma^2}{M}+\gamma^2\sigma^2\end{aligned}$$

Now we sum up both inequalities and we get:

$$\begin{aligned}&\mathbb{E}[\Psi(\mathbf{z}^{m,r}_{k+1}) + \Phi(\bar{\mathbf{z}}^r_{k+1})]\\ &\leqslant \left(1+\frac{1}{K}\right)\mathbb{E}[\Psi(\mathbf{z}^{m,r}_k) + \Phi(\bar{\mathbf{z}}^r_k)] + 8KL^2\gamma^2\,\mathbb{E}[\Psi(\mathbf{z}^{m,r}_k)+\Phi(\bar{\mathbf{z}}^r_k)] + \\ &\qquad 32KL^2\gamma^2\,\mathbb{E}\left\|\bar{\mathbf{z}}^r_k - \mathbf{z}^\star\right\|^2 + 32K\gamma^2\zeta_\star^2 + \frac{2\gamma^2\sigma^2}{M}+2\gamma^2\sigma^2\end{aligned}$$

With the choice of $\gamma \leqslant \frac{\mu}{32L^2K}$ we simplify the above inequality as:

$$\begin{aligned}&\mathbb{E}[\Psi(\mathbf{z}^{m,r}_{k+1}) + \Phi(\bar{\mathbf{z}}^r_{k+1})]\\ &\leqslant \left(1+\frac{1}{K}+\frac{1}{128K}\right)\mathbb{E}[\Psi(\mathbf{z}^{m,r}_k) + \Phi(\bar{\mathbf{z}}^r_k)] + \frac{\mu^2}{32KL^2}\,\mathbb{E}\left\|\bar{\mathbf{z}}^r_k - \mathbf{z}^\star\right\|^2 + 32K\gamma^2\zeta_\star^2 + \frac{2\gamma^2\sigma^2}{M}+2\gamma^2\sigma^2\end{aligned}$$

After unrolling the recursion for the last K steps and considering the fact that $\left(1+\frac{1}{K}+\frac{1}{128K}\right)^K \leqslant 4$ we have:

$$\mathbb{E}[\Psi(\mathbf{z}^{m,r}_{k+1}) + \Phi(\bar{\mathbf{z}}^r_{k+1})] \leqslant \sum_{i=1}^{K}\frac{\mu^2}{8KL^2}\,\mathbb{E}\left\|\bar{\mathbf{z}}^r_i - \mathbf{z}^\star\right\|^2 + 32K^2\gamma^2\zeta_\star^2 + \frac{2K\gamma^2\sigma^2}{M}+2K\gamma^2\sigma^2$$

□

Lemma 14 (Consensus error for N-player games). *After running Decoupled SGD for k local steps at some round r with a step-size of $\gamma \leqslant \frac{\mu}{32L^2K}$, the error $\sum_{n=1}^{N}\Phi(\mathbf{x}^{m,r}_k)$ can be upper bounded as follows:*

$$\sum_{n=1}^{N}\mathbb{E}[\Phi(\mathbf{x}^{n,r}_k)] \leqslant \sum_{i=1}^{K}\frac{\mu^2}{64KNL^2}\,\mathbb{E}\left\|\mathbf{z}^r_i - \mathbf{z}^\star\right\|^2 + 4NK\gamma^2\sigma^2 \tag{24}$$

Proof. We start by upper bounding this error for some player n:

$$\begin{aligned}&\mathbb{E}[\Phi(\mathbf{x}^{n,r}_{k+1})]\\ &= \mathbb{E}\left\|\mathbf{x}^{n,r}_{k+1} - \mathbf{x}^{n,r}_0\right\|^2\\ &= \mathbb{E}\left\|\mathbf{x}^{n,r}_k - \nabla_{\mathbf{x}^n} f_n(\mathbf{x}^{n,r}_k;\mathbf{x}^{-n,r}_0;\xi_n) - \mathbf{x}^{n,r}_0\right\|^2\\ &\leqslant \mathbb{E}\left\|\mathbf{x}^{n,r}_k - \nabla_{\mathbf{x}^n} f_m(\mathbf{x}^{n,r}_k;\mathbf{x}^{-n,r}_0) - \mathbf{x}^{n,r}_0\right\|^2 + \gamma^2\sigma^2\\ &\leqslant \left(1+\frac{1}{K}\right)\mathbb{E}[\Phi(\mathbf{x}^{n,r}_k)] + 2K\gamma^2\,\mathbb{E}\left\|\nabla_{\mathbf{x}^n} f_n(\mathbf{x}^{n,r}_k;\mathbf{x}^{-n,r}_0)\right\|^2 + \gamma^2\sigma^2\\ &\leqslant \left(1+\frac{1}{K}\right)\mathbb{E}[\Phi(\mathbf{x}^{n,r}_k)] + 2K\gamma^2\,\mathbb{E}\left\|\nabla_{\mathbf{x}^n} f_n(\mathbf{x}^{n,r}_k;\mathbf{x}^{-n,r}_0) - \nabla_{\mathbf{x}^n} f_n(\mathbf{x}^{n,r}_k,\mathbf{x}^{-n,r}_k) + \nabla_{\mathbf{x}^n} f_n(\mathbf{x}^{n,r}_k,\mathbf{x}^{-n,r}_k)\right\|^2 + \gamma^2\sigma^2\\ &\leqslant \left(1+\frac{1}{K}\right)\mathbb{E}[\Phi(\mathbf{x}^{n,r}_k)] + 4KL^2\gamma^2\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{x}^{i,r}_k - \mathbf{x}^{i,r}_0\right\|^2 + 4KL^2\gamma^2\,\mathbb{E}\left\|\mathbf{z}^r_k - \mathbf{z}^\star\right\|^2 + \gamma^2\sigma^2\end{aligned}$$

Summing over all players gives us:

$$\begin{aligned}
&\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k+1}^{n,r})] \\
&\leqslant \left(1+\frac{1}{K}\right)\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k}^{n,r})] + 4KL^2\gamma^2 \sum_{n=1}^{N}\sum_{i=1}^{N} \mathbb{E}\left\|\mathbf{x}_k^{i,r} - \mathbf{x}_0^{i,r}\right\|^2 + 4KNL^2\gamma^2\,\mathbb{E}\left\|\mathbf{z}_k^r - \mathbf{z}^\star\right\|^2 + N\gamma^2\sigma^2 \\
&\leqslant \left(1+\frac{1}{K}\right)\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k}^{n,r})] + 4KNL^2\gamma^2 \sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k}^{n,r})] + 4KNL^2\gamma^2\,\mathbb{E}\left\|\mathbf{z}_k^r - \mathbf{z}^\star\right\|^2 + N\gamma^2\sigma^2
\end{aligned}$$

With the choice of $\gamma \leqslant \frac{\mu}{32NKL^2}$ we get:

$$\begin{aligned}
&\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k+1}^{n,r})] \\
&\leqslant \left(1+\frac{1}{K}\right)\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k}^{n,r})] + \frac{\mu^2}{256KNL^2}\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k}^{n,r})] + \frac{\mu^2}{256KNL^2}\,\mathbb{E}\left\|\mathbf{z}_k^r - \mathbf{z}^\star\right\|^2 + N\gamma^2\sigma^2 \\
&= \left(1+\frac{1}{K}+\frac{\mu^2}{256KNL^2}\right)\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k}^{n,r})] + \frac{\mu^2}{256KNL^2}\,\mathbb{E}\left\|\mathbf{z}_k^r - \mathbf{z}^\star\right\|^2 + N\gamma^2\sigma^2
\end{aligned}$$

By unrolling the recursion for K steps and considering the fact that $\left(1+\frac{1}{K}+\frac{\mu^2}{256KNL^2}\right)^K \leqslant 4$ we get:

$$\sum_{n=1}^{N} \mathbb{E}[\Phi(\mathbf{x}_{k+1}^{n,r})] \leqslant \sum_{i=1}^{K} \frac{\mu^2}{64KNL^2}\,\mathbb{E}\left\|\mathbf{z}_i^r - \mathbf{z}^\star\right\|^2 + 4NK\gamma^2\sigma^2$$

□

A.2 Proof of Theorem 4

We begin by upper bounding the distance between the average iterate $\bar{\mathbf{x}}_{k+1}^r$ and the saddle point.

Proof.

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{x}}_{k+1}^r - \mathbf{x}^\star\right\|^2 \\
&= \mathbb{E}\left\|\bar{\mathbf{x}}_k^r - \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r};\xi_m) - \mathbf{x}^\star\right\|^2 \\
&\leqslant \mathbb{E}\left\|\bar{\mathbf{x}}_k^r - \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) - \mathbf{x}^\star\right\|^2 + \frac{\gamma^2\sigma^2}{M} \\
&= \mathbb{E}\left\|\bar{\mathbf{x}}_k^r + \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^r, \bar{\mathbf{y}}_k^r) - \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r}) - \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^r, \bar{\mathbf{y}}_k^r) - \mathbf{x}^\star\right\|^2 + \frac{\gamma^2\sigma^2}{M} \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left\|\bar{\mathbf{x}}_k^r - \frac{\gamma}{M}\sum_{m=1}^{M} \nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^r, \bar{\mathbf{y}}_k^r) - \mathbf{x}^\star\right\|^2 + \\
&\qquad\qquad \left(1+\frac{2}{\gamma\mu}\right)\frac{\gamma^2}{M}\sum_{m=1}^{M} \mathbb{E}\left\|\nabla_{\mathbf{x}} f_m(\bar{\mathbf{x}}_k^r, \bar{\mathbf{y}}_k^r) - \nabla_{\mathbf{x}} f_m(\mathbf{x}_k^{m,r}, \mathbf{y}_0^{m,r})\right\|^2 + \frac{\gamma^2\sigma^2}{M}
\end{aligned}$$

For the first term in the above inequality we have:

$$\begin{aligned}
&\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\left\|\bar{\mathbf{x}}_k^r-\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{x}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\mathbf{x}^\star\right\|^2\\
&=\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\left\|\bar{\mathbf{x}}_k^r-\gamma\nabla_{\mathbf{x}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\mathbf{x}^\star\right\|^2\\
&=\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\Big[\left\|\bar{\mathbf{x}}_k^r-\mathbf{x}^\star\right\|^2+\gamma^2\left\|\nabla_{\mathbf{x}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)\right\|^2-2\gamma\langle\bar{\mathbf{x}}_k^r-\mathbf{x}^\star,\nabla_{\mathbf{x}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)\rangle\Big]\\
&\leqslant\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\Big[(1+\gamma^2L^2)\left\|\bar{\mathbf{x}}_k^r-\mathbf{x}^\star\right\|^2-2\gamma\langle\bar{\mathbf{x}}_k^r-\mathbf{x}^\star,\nabla_{\mathbf{x}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)\rangle\Big]
\end{aligned}$$

For the second term we also have:

$$\begin{aligned}
&\left(1+\frac{2}{\gamma\mu}\right)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\nabla_{\mathbf{x}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\nabla_{\mathbf{x}}f_m(\mathbf{x}_k^{m,r},\mathbf{y}_0^{m,r})\right\|^2\\
&\leqslant\left(1+\frac{2}{\gamma\mu}\right)\frac{L^2\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\bar{\mathbf{x}}_k^r-\mathbf{x}_k^{m,r}\right\|^2+\left(1+\frac{2}{\gamma\mu}\right)\frac{L^2\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\bar{\mathbf{y}}_k^r-\bar{\mathbf{y}}_0^r\right\|^2\\
&=\left(1+\frac{2}{\gamma\mu}\right)L^2\gamma^2\,\mathbb{E}\left[\Psi(\mathbf{x}_k^{m,r})\right]+\left(1+\frac{2}{\gamma\mu}\right)L^2\gamma^2\,\mathbb{E}[\Phi(\bar{\mathbf{y}}_k^r)]
\end{aligned}$$

Where in the last line, we used the fact that $\mathbf{y}_0^{m,r}=\bar{\mathbf{y}}_0^r$. We then repeat the same computation with respect to $\mathbf{y}$.

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{y}}_{k+1}^r-\mathbf{y}^\star\right\|^2=\\
&=\mathbb{E}\left\|\bar{\mathbf{y}}_k^r+\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\mathbf{x}_0^{m,r},\mathbf{y}_k^{m,r};\xi_m)-\mathbf{y}^\star\right\|^2\\
&\leqslant\mathbb{E}\left\|\bar{\mathbf{y}}_k^r+\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\mathbf{x}_0^{m,r},\mathbf{y}_k^{m,r})-\mathbf{y}^\star\right\|^2+\frac{\gamma\sigma^2}{M}\\
&=\mathbb{E}\left\|\bar{\mathbf{y}}_k^r+\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\mathbf{x}_0^{m,r},\mathbf{y}_k^{m,r})-\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)+\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\mathbf{y}^\star\right\|^2+\frac{\gamma\sigma^2}{M}\\
&\leqslant\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\left\|\bar{\mathbf{y}}_k^r+\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\mathbf{y}^\star\right\|^2+\\
&\qquad\qquad\left(1+\frac{2}{\gamma\mu}\right)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\nabla_{\mathbf{y}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\nabla_{\mathbf{y}}f_m(\mathbf{x}_0^{m,r},\mathbf{y}_k^{m,r})\right\|^2+\frac{\gamma\sigma^2}{M}
\end{aligned}$$

For the first term in the above inequality we have:

$$\begin{aligned}
&\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\left\|\bar{\mathbf{y}}_k^r+\frac{\gamma}{M}\sum_{m=1}^{M}\nabla_{\mathbf{y}}f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\mathbf{y}^\star\right\|^2\\
&=\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\left\|\bar{\mathbf{y}}_k^r+\gamma\nabla_{\mathbf{y}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\mathbf{y}^\star\right\|^2\\
&=\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\Big[\left\|\bar{\mathbf{y}}_k^r-\mathbf{y}^\star\right\|^2+\gamma^2\left\|\nabla_{\mathbf{y}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)\right\|^2-2\gamma\langle\mathbf{y}^\star-\bar{\mathbf{y}}_k^r,\nabla_{\mathbf{y}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)\rangle\Big]\\
&\leqslant\Big(1+\frac{\gamma\mu}{2}\Big)\,\mathbb{E}\Big[(1+\gamma^2L^2)\left\|\bar{\mathbf{y}}_k^r-\mathbf{y}^\star\right\|^2-2\gamma\langle\mathbf{y}^\star-\bar{\mathbf{y}}_k^r,\nabla_{\mathbf{y}}f(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)\rangle\Big]
\end{aligned}$$

For the second term we also have:

$$
\begin{aligned}
&\left(1+\frac{2}{\gamma\mu}\right)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\nabla_{\mathbf{y}} f_m(\bar{\mathbf{x}}_k^r,\bar{\mathbf{y}}_k^r)-\nabla_{\mathbf{y}} f_m(\mathbf{x}_0^{m,r},\mathbf{y}_k^{m,r})\right\|^2\\
&\leqslant \left(1+\frac{2}{\gamma\mu}\right)\frac{L^2\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\bar{\mathbf{x}}_k^r-\bar{\mathbf{x}}_0^r\right\|^2+\left(1+\frac{2}{\gamma\mu}\right)\frac{L^2\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\bar{\mathbf{y}}_k^r-\mathbf{y}_k^{m,r}\right\|^2\\
&=\left(1+\frac{2}{\gamma\mu}\right)L^2\gamma^2\,\mathbb{E}[\Phi(\bar{\mathbf{x}}_k^r)]+\left(1+\frac{2}{\gamma\mu}\right)L^2\gamma^2\,\mathbb{E}[\Psi(\mathbf{y}_k^{m,r})]
\end{aligned}
$$

Summing up the results from the inequalities with respect to $\mathbf{x}$ and $\mathbf{y}$ gives us:

$$
\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{z}}_{k+1}^r-\mathbf{z}^\star\right\|^2\\
&\leqslant\left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[(1+\gamma^2L^2)\left\|\bar{\mathbf{z}}_k^r-\mathbf{z}^\star\right\|^2-2\gamma\langle\bar{\mathbf{z}}_k^r-\mathbf{z}^\star,F(\bar{\mathbf{z}}_k^r)\rangle\right]+\gamma\left(\gamma L^2+\frac{2L^2}{\mu}\right)\mathbb{E}\left[\Phi(\bar{\mathbf{z}}_k^r)+\Psi(\mathbf{z}_k^{m,r})\right]+\frac{\gamma^2\sigma^2}{M}\\
&\leqslant\left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[(1+\gamma^2L^2)\left\|\bar{\mathbf{z}}_k^r-\mathbf{z}^\star\right\|^2-2\gamma\mu\left\|\bar{\mathbf{z}}_k^r-\mathbf{z}^\star\right\|^2\right]+\gamma\left(\gamma L^2+\frac{2L^2}{\mu}\right)\mathbb{E}\left[\Phi(\bar{\mathbf{z}}_k^r)+\Psi(\mathbf{z}_k^{m,r})\right]+\frac{\gamma^2\sigma^2}{M}\\
&=\left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[(1-2\gamma\mu+\gamma^2L^2)\left\|\bar{\mathbf{z}}_k^r-\mathbf{z}^\star\right\|^2\right]+\gamma\left(\gamma L^2+\frac{2L^2}{\mu}\right)\mathbb{E}\left[\Phi(\bar{\mathbf{z}}_k^r)+\Psi(\mathbf{z}_k^{m,r})\right]+\frac{\gamma^2\sigma^2}{M}
\end{aligned}
$$

With the choice of $\gamma\leqslant\frac{\mu}{16L^2}$ we have:

$$
\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{z}}_{k+1}^r-\mathbf{z}^\star\right\|^2\\
&\leqslant\left(1-\frac{23\gamma\mu}{16}\right)\mathbb{E}\left\|\bar{\mathbf{z}}_k^r-\mathbf{z}^\star\right\|^2+\frac{33\gamma L^2}{16\mu}\mathbb{E}\left[\Phi(\bar{\mathbf{z}}_k^r)+\Psi(\mathbf{z}_k^{m,r})\right]+\frac{\gamma^2\sigma^2}{M}\\
&\leqslant\left(1-\frac{23\gamma\mu}{16}\right)\mathbb{E}\left\|\bar{\mathbf{z}}_k^r-\mathbf{z}^\star\right\|^2+\frac{33\gamma\mu}{128K}\sum_{i=1}^{K}\left\|\bar{\mathbf{z}}_i^r-\mathbf{z}^\star\right\|^2+\frac{96K^2L^2\gamma^3\zeta_\star^2}{\mu}+\frac{7KL^2\gamma^3\sigma^2}{\mu M}+\frac{6KL^2\gamma^3\sigma^2}{\mu}+\frac{\gamma^2\sigma^2}{M}
\end{aligned}
$$

We change the current notation for simplicity in proof by substituting r and k with t. t varies from 0 to $T=KR$, iterating over all rounds and local steps:

$$
\begin{aligned}
\mathbb{E}\left\|\bar{\mathbf{z}}_{t+1}-\mathbf{z}^\star\right\|^2\leqslant\left(1-\frac{23\gamma\mu}{16}\right)\mathbb{E}\left\|\bar{\mathbf{z}}_t-\mathbf{z}^\star\right\|^2+\frac{33\gamma\mu}{128K}\sum_{i=\max\{0,t-K+1\}}^{t}\left\|\bar{\mathbf{z}}_i-\mathbf{z}^\star\right\|^2\\
+\frac{96K^2L^2\gamma^3\zeta_\star^2}{\mu}+\frac{7KL^2\gamma^3\sigma^2}{\mu M}+\frac{6KL^2\gamma^3\sigma^2}{\mu}+\frac{\gamma^2\sigma^2}{M}
\end{aligned}
$$

Here we use the Lemma 11 with the following parameters,

$$
s_t=\mathbb{E}\left\|\bar{\mathbf{z}}_t-\mathbf{z}^\star\right\|^2\ ,\ a=\frac{23\mu}{16}\ ,\ b=\frac{33\mu}{128}\ ,\ c=\frac{96K^2L^2\gamma\zeta_\star^2}{\mu}+\frac{7KL^2\gamma\sigma^2}{\mu M}+\frac{6KL^2\gamma\sigma^2}{\mu}+\frac{\gamma\sigma^2}{M}
$$

The final inequality is:

$$
\begin{aligned}
\mathbb{E}\left\|\bar{\mathbf{z}}_t-\mathbf{z}^\star\right\|^2&\leqslant\left(1-\frac{23\gamma\mu}{32}\right)^t\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2+\frac{32}{23\mu}\left(\frac{96K^2L^2\gamma\zeta_\star^2}{\mu}+\frac{7KL^2\gamma\sigma^2}{\mu M}+\frac{6KL^2\gamma\sigma^2}{\mu}+\frac{\gamma\sigma^2}{M}\right)\gamma\\
&\leqslant\left(1-\frac{\gamma\mu}{2}\right)^t\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2+\frac{96K^2L^2\gamma^2\zeta_\star^2}{\mu^2}+\frac{7KL^2\gamma^2\sigma^2}{\mu^2M}+\frac{6KL^2\gamma^2\sigma^2}{\mu^2}+\frac{\gamma\sigma^2}{M\mu}
\end{aligned}
$$

Recall that we assumed $\gamma=\frac{\mu}{32KL^2}$.By using this inequality we can dirve :

$$
\mathbb{E}\left\|\bar{\mathbf{z}}_T-\mathbf{z}^\star\right\|^2\leqslant\left(1-\frac{\gamma\mu}{2}\right)^{KR}\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2+\frac{96K^2L^2\gamma^2\zeta_\star^2}{\mu^2}+\frac{6KL^2\gamma^2\sigma^2}{\mu^2}+\frac{2\gamma\sigma^2}{M\mu}
$$

By setting $t=T=RK$, we get:

$$
\begin{aligned}
\mathbb{E}\left\|\bar{\mathbf{z}}_T-\mathbf{z}^\star\right\|^2&\leqslant\left(1-\frac{\gamma\mu}{2}\right)^{KR}\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2+\frac{96K^2L^2\gamma^2\zeta_\star^2}{\mu^2}+\frac{6KL^2\gamma^2\sigma^2}{\mu^2}+\frac{2\gamma\sigma^2}{M\mu}\\
&\leqslant\exp\left(-\frac{\gamma\mu}{2}KR\right)\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2+\frac{96K^2L^2\gamma^2\zeta_\star^2}{\mu^2}+\frac{6KL^2\gamma^2\sigma^2}{\mu^2}+\frac{2\gamma\sigma^2}{M\mu}
\end{aligned}
$$

We can see that with this inequality we can only guarantee convergence to a neighborhood of $\mathbf{z}^\star$. To obtain a convergence the final, as discussed in Stich [2019], we need to choose the step size carefully. If $\frac{\mu}{32KL^2} \geqslant \frac{\ln(\max\{2,\mu^4\|\mathbf{z}_0-\mathbf{z}^\star\|^2T^2/\sigma^2\})}{\mu T}$ then we choose $\gamma = \frac{\ln(\max\{2,\mu^4\|\mathbf{z}_0-\mathbf{z}^\star\|^2T^2/\sigma^2\})}{\mu T}$, otherwise if $\frac{\mu}{32KL^2} < \frac{\ln(\max\{2,\mu^4\|\mathbf{z}_0-\mathbf{z}^\star\|^2T^2/\sigma^2\})}{\mu T}$ then we choose $\gamma = \frac{\mu}{32KL^2}$

we can see that with these choices, we would have:

$$\mathbb{E}\left\|\bar{\mathbf{z}}_T - \mathbf{z}^\star\right\|^2 = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu^2}{64L^2}R\right)\|\mathbf{z}_0-\mathbf{z}^\star\|^2 + \frac{K^2L^2\zeta_\star^2}{\mu^4T^2} + \frac{KL^2\sigma^2}{\mu^4T^2} + \frac{2\sigma^2}{M\mu^2T}\right)$$

□

A.3 Proof of Theorem 1

Proof.

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{x}_{k+1}^{n,r} - \mathbf{x}_\star^n\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n - \gamma\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_0^{-n,r};\xi_n)\right\|^2 \\
&\leqslant \mathbb{E}\left\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n - \gamma\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_0^{-n,r})\right\|^2 + \gamma^2\sigma^2 \\
&= \mathbb{E}\left\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n - \gamma\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r}) - \gamma\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_0^{-n,r}) + \gamma\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\|^2 + \gamma^2\sigma^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n - \gamma\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\|^2 \\
&\qquad + \left(1+\frac{2}{\gamma\mu}\right)\gamma^2\mathbb{E}\left\|\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_0^{-n,r}) - \nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\|^2 + \gamma^2\sigma^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n\|^2 + \gamma^2\left\|\nabla_{\mathbf{x}^n} f(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\|^2 - 2\gamma\left\langle\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n, \nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\rangle\right] \\
&\qquad + \left(\gamma^2+\frac{2\gamma}{\mu}\right)\mathbb{E}\left\|\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r}) - \nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_0^{-n,r})\right\|^2 + \gamma^2\sigma^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n\|^2 + \gamma^2\left\|\nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\|^2 - 2\gamma\left\langle\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n, \nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\rangle\right] \\
&\qquad + \left(\gamma^2L^2+\frac{2\gamma L^2}{\mu}\right)\sum_{i=1}^N\mathbb{E}\left\|\mathbf{x}_k^{i,r} - \mathbf{x}_0^{i,r}\right\|^2 + \gamma^2\sigma^2
\end{aligned}$$

Then we sum up both sides of the above inequality over n:

$$\begin{aligned}
&\sum_{n=1}^N\mathbb{E}\left\|\mathbf{x}_{k+1}^{n,r} - \mathbf{x}_\star^n\right\|^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[(1+\gamma^2L^2)\sum_{n=1}^N\|\mathbf{x}_k^{n,r} - \mathbf{x}_\star^n\|^2 - 2\gamma\sum_{n=1}^N\left\langle\mathbf{x}_k^{n,r} - \mathbf{x}_\star^N, \nabla_{\mathbf{x}^n} f_n(\mathbf{x}_k^{n,r};\mathbf{x}_k^{-n,r})\right\rangle\right] \\
&\qquad + \left(\gamma^2L^2+\frac{2\gamma L^2}{\mu}\right)N\sum_{n=1}^N\mathbb{E}\|\mathbf{x}_k^{n,r} - \mathbf{x}_0^{n,r}\|^2 + \gamma^2\sigma^2 \\
&= \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[(1+\gamma^2L^2)\|\mathbf{z}_k^r - \mathbf{z}_\star\|^2 - 2\gamma\left\langle\mathbf{z}_k^r - \mathbf{z}_\star, F(\mathbf{z}_k^r) - F(\mathbf{z}_\star)\right\rangle\right] + \left(\gamma^2L^2+\frac{2\gamma L^2}{\mu}\right)N\Phi(\mathbf{x}_k^{n,r}) + \gamma^2\sigma^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\mathbb{E}\left[(1-2\gamma\mu+\gamma^2L^2)\|\mathbf{z}_k^r - \mathbf{z}_\star\|^2\right] + \left(\gamma^2L^2+\frac{2\gamma L^2}{\mu}\right)N\Phi(\mathbf{x}_k^{n,r}) + \gamma^2\sigma^2
\end{aligned}$$

With the choice of $\gamma \leqslant \frac{\mu}{8L^2}$ we have:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{z}_{k+1}^r - \mathbf{z}^\star\right\|^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\left[\left(1-2\gamma\mu+\gamma^2L^2\right)\mathbb{E}\left\|\mathbf{z}_k^r-\mathbf{z}^\star\right\|^2\right] + \gamma\left(\gamma L^2+\frac{4L^2}{\mu}\right)N\Phi(\mathbf{x}_k^{n,r}) + \gamma^2\sigma^2 \\
&\leqslant \left(1+\frac{\gamma\mu}{2}\right)\left[\left(1-\frac{15\gamma\mu}{8}\right)\mathbb{E}\left\|\mathbf{z}_k^r-\mathbf{z}^\star\right\|^2\right] + \frac{33\gamma L^2 N}{8\mu}\Phi(\mathbf{x}_k^{n,r}) + \gamma^2\sigma^2 \\
&\leqslant \left(1-\frac{11\gamma\mu}{8}\right)\mathbb{E}\left\|\mathbf{z}_k^r-\mathbf{z}^\star\right\|^2 + \frac{33\gamma L^2 N}{8\mu}\Phi(\mathbf{x}_k^{n,r}) + \gamma^2\sigma^2 \\
&\leqslant \left(1-\frac{11\gamma\mu}{8}\right)\mathbb{E}\left\|\mathbf{z}_k^r-\mathbf{z}^\star\right\|^2 + \frac{33\gamma\mu}{512K}\sum_{i=1}^{K}\mathbb{E}\left\|\mathbf{z}_i^r-\mathbf{z}^\star\right\|^2 + \gamma^2\sigma^2 + \frac{20NKL^2\gamma^3\sigma^2}{\mu}
\end{aligned}$$

We change the current notation for simplicity in proof by substituting r and k with t. t varies from 0 to $T = KR$, iterating over all rounds and local steps:

$$\mathbb{E}\left\|\mathbf{z}_{t+1}-\mathbf{z}^\star\right\|^2 \leqslant \left(1-\frac{11\gamma\mu}{8}\right)\mathbb{E}\left\|\mathbf{z}_t-\mathbf{z}^\star\right\|^2 + \frac{33\gamma\mu}{512K}\sum_{i=\max\{0,t-K+1\}}^{t}\mathbb{E}\left\|\mathbf{z}_i-\mathbf{z}^\star\right\|^2 + \gamma^2\sigma^2 + \frac{20NKL^2\gamma^3\sigma^2}{\mu}$$

Here we use the Lemma 11 with the following parameters,

$$s_t = \mathbb{E}\left\|\mathbf{z}_t-\mathbf{z}^\star\right\|^2 \ , \ a = \frac{11\mu}{8} \ , \ b = \frac{33\mu}{512} \ , \ c = \left(1+\frac{20NKL^2\gamma}{\mu}\right)\sigma^2$$

The final inequality is:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}_t-\mathbf{z}^\star\right\|^2 &\leqslant \left(1-\frac{11\gamma\mu}{16}\right)^t\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2 + \frac{16}{11\mu}\gamma\sigma^2 + \frac{320NKL^2\gamma^2\sigma^2}{11\mu^2} \\
&\leqslant \left(1-\frac{\gamma\mu}{2}\right)^t\mathbb{E}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2 + \frac{2}{\mu}\gamma\sigma^2 + \frac{30NKL^2\gamma^2\sigma^2}{\mu^2}
\end{aligned}$$

By setting $t = T$ and by considering the inequality $\gamma \leqslant \frac{\mu}{32NKL^2}$, we get:

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}_T-\mathbf{z}^\star\right\|^2 &\leqslant \left(1-\frac{\gamma\mu}{2}\right)^{KR}\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2 + \frac{2}{\mu}\gamma\sigma^2 + \frac{1}{\mu}\gamma\sigma^2 \\
&\leqslant \exp\left(-\frac{\gamma\mu}{2}KR\right)\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2 + \frac{3}{\mu}\gamma\sigma^2
\end{aligned}$$

We can see that with this inqualty we can only guarantee convergence to a $\frac{\gamma\sigma^2}{\mu}$-neighborhood of $\mathbf{z}^\star$. To obtain a convergence the final, as discussed in Stich [2019], we need to choose the step size carefully. If $\frac{\mu}{32NKL^2} \geqslant \frac{\ln(\max\{2,\mu^2\|\mathbf{z}_0-\mathbf{z}^\star\|^2T/\sigma^2\})}{\mu T}$ then we choose $\gamma = \frac{\ln(\max\{2,\mu^2\|\mathbf{z}_0-\mathbf{z}^\star\|^2T/\sigma^2\})}{\mu T}$,otherwise if $\frac{\mu}{32NKL^2} < \frac{\ln(\max\{2,\mu^2\|\mathbf{z}_0-\mathbf{z}^\star\|^2T/\sigma^2\})}{\mu T}$ then we choose $\gamma = \frac{\mu}{32NKL^2}$

we can see that with these choices, we would have:

$$\mathbb{E}\left\|\mathbf{z}_T-\mathbf{z}^\star\right\|^2 = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu^2}{64NL^2}R\right)\left\|\mathbf{z}_0-\mathbf{z}^\star\right\|^2 + \frac{\sigma^2}{\mu^2KR}\right)$$

□

Lemma 15. *Given a general bi-linear game in the following form:*

$$f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \frac{1}{2}\mathbf{y}^\top\mathbf{B}\mathbf{y} + \mathbf{x}^\top\mathbf{C}\mathbf{y}$$

After k steps of Decoupled GDA at some round r we can compute the explicit form of iterates as follows:

$$\mathbf{x}_k^r = -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I}-\gamma\mathbf{A}\right)^k\left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right)$$

$$\mathbf{y}_k^r = \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I}-\gamma\mathbf{B}\right)^k\left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)$$

Proof. We use induction for the proof of this section. By using the update rule of Local GDA we would have,

$$\begin{aligned}
\mathbf{x}_{k+1}^r &= \mathbf{x}_k - \gamma \nabla_{\mathbf{x}} f(\mathbf{x}_k^r, \mathbf{y}_0^r) \\
&= \mathbf{x}_k - \gamma \left(\mathbf{A}\mathbf{x}_k^r + \mathbf{C}\mathbf{y}_0^r\right) \\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right) \\
&\qquad - \gamma \Big(\mathbf{A}\Big[-\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0\right)\Big] + \mathbf{C}\mathbf{y}_0^r\Big) \\
\\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right) \\
&\qquad - \gamma \Big(-\mathbf{C}\mathbf{y}_0^r + \left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right) + \mathbf{C}\mathbf{y}_0^r\Big) \\
\\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right) - \gamma\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right) \\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \left(\mathbf{A}^{-1} - \gamma\mathbf{I}\right)\Big[\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right)\Big] \\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)\Big[\left(\mathbf{I} - \gamma\mathbf{A}\right)^k \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right)\Big] \\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)^{k+1} \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right)
\end{aligned}$$

Now we only need to show that our claim also works for $k = 0$,

$$\begin{aligned}
\mathbf{x}_0^r &= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I} - \gamma\mathbf{A}\right)^0 \left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right) \\
&= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{x}_0^r + \mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r \\
&= \mathbf{x}_0^r
\end{aligned}$$

Also, we do the computation with respect to $\mathbf{y}$:

$$\mathbf{y}_k^r = \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)$$

By using the update rule of Local GDA we get:

$$\begin{aligned}
\mathbf{y}_{k+1}^r &= \mathbf{y}_k - \gamma \nabla_{\mathbf{y}} f(\mathbf{x}_0^r, \mathbf{x}_k^r) \\
&= \mathbf{y}_k + \gamma \left(-\mathbf{B}\mathbf{y}_k^r + \mathbf{C}^\top\mathbf{x}_0^r\right) \\
&= \mathbf{y}_k - \gamma \left(\mathbf{B}\mathbf{y}_k^r - \mathbf{C}^\top\mathbf{x}_0^r\right) \\
&= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0\right) \\
&\qquad - \gamma \Big(\mathbf{B}\Big[\mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)\Big] - \mathbf{C}^\top\mathbf{x}_0^r\Big) \\
\\
&= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0\right) \\
&\qquad - \gamma \Big(\mathbf{C}^\top\mathbf{x}_0^r + \left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right) - \mathbf{C}^\top\mathbf{x}_0^r\Big) \\
\\
&= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0\right) - \gamma\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right) \\
&= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \left(\mathbf{B}^{-1} - \gamma\mathbf{I}\right)\Big[\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)\Big] \\
&= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)\Big[\left(\mathbf{I} - \gamma\mathbf{B}\right)^k \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)\Big] \\
&= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I} - \gamma\mathbf{B}\right)^{k+1} \left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)
\end{aligned}$$

Now we only need to show this our claim also works for $k = 0$,

$$\begin{aligned}\mathbf{y}_0^r &= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I}-\gamma\mathbf{B}\right)^0\left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)\\ &= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{y}_0^r - \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r\\ &= \mathbf{y}_0^r\end{aligned}$$

□

A.4 Proof of Theorem 1

Proof. Here we provide the convergence rate for Decoupled GDA for bi-linear games in the form of:

Recall that from Lemma 15 we know:

$$\begin{aligned}\mathbf{x}_k^r &= -\mathbf{A}^{-1}\mathbf{C}\mathbf{y}_0^r + \mathbf{A}^{-1}\left(\mathbf{I}-\gamma\mathbf{A}\right)^k\left(\mathbf{A}\mathbf{x}_0^r + \mathbf{C}\mathbf{y}_0^r\right)\\ \mathbf{y}_k^r &= \mathbf{B}^{-1}\mathbf{C}^\top\mathbf{x}_0^r + \mathbf{B}^{-1}\left(\mathbf{I}-\gamma\mathbf{B}\right)^k\left(\mathbf{B}\mathbf{y}_0^r - \mathbf{C}^\top\mathbf{x}_0^r\right)\end{aligned}$$

Then we can write our expressions in the matrix form:

$$\begin{aligned}\mathbf{z}_k^r &= \begin{pmatrix}\mathbf{A}^{-1}\left(\mathbf{I}-\gamma\mathbf{A}\right)^k\mathbf{A} & \mathbf{A}^{-1}\left(-\mathbf{C}+\left(\mathbf{I}-\gamma\mathbf{A}\right)^k\mathbf{C}\right)\\ \mathbf{B}^{-1}\left(\mathbf{C}^\top - \left(\mathbf{I}-\gamma\mathbf{B}\right)^k\mathbf{C}^\top\right) & \mathbf{B}^{-1}\left(\mathbf{I}-\gamma\mathbf{B}\right)^k\mathbf{B}\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\\[2ex] &= \begin{pmatrix}\left(\mathbf{I}-\gamma\mathbf{A}\right)^k & \mathbf{A}^{-1}\left(-\mathbf{C}+\left(\mathbf{I}-\gamma\mathbf{A}\right)^k\mathbf{C}\right)\\ \mathbf{B}^{-1}\left(\mathbf{C}^\top - \left(\mathbf{I}\gamma\mathbf{B}\right)^k\mathbf{C}^\top\right) & \left(\mathbf{I}-\gamma\mathbf{B}\right)^k\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\\[2ex] &= \left[\begin{pmatrix}\left(\mathbf{I}-\gamma\mathbf{A}\right)^k & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{I}-\gamma\mathbf{B}\right)^k\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{A}^{-1}\mathbf{C}\\ -\mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{I}\end{pmatrix} + \begin{pmatrix}\mathbf{0} & -\mathbf{A}^{-1}\mathbf{C}\\ \mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{0}\end{pmatrix}\right]\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\\[2ex] &= \left[\begin{pmatrix}\left(\mathbf{I}-\gamma\mathbf{A}\right) & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{I}-\gamma\mathbf{B}\right)\end{pmatrix}^k\begin{pmatrix}\mathbf{I} & \mathbf{A}^{-1}\mathbf{C}\\ -\mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{I}\end{pmatrix} + \begin{pmatrix}\mathbf{0} & -\mathbf{A}^{-1}\mathbf{C}\\ \mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{0}\end{pmatrix}\right]\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\end{aligned}$$

Then we compute the norm squared of $\|\mathbf{z}_k^r\|^2$:

$$\|\mathbf{z}_k^r\|^2 = \left\|\left[\begin{pmatrix}\left(\mathbf{I}-\gamma\mathbf{A}\right) & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{I}-\gamma\mathbf{B}\right)\end{pmatrix}^k\begin{pmatrix}\mathbf{I} & \mathbf{A}^{-1}\mathbf{C}\\ -\mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{I}\end{pmatrix} + \begin{pmatrix}\mathbf{0} & -\mathbf{A}^{-1}\mathbf{C}\\ \mathbf{B}^{-1}\mathbf{C}^\top & \mathbf{0}\end{pmatrix}\right]\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\right\|^2$$

Recall that for simplifying the proof, here we study the bi-linear games where $\mathbf{A} = \omega\mathbf{I}$, $\mathbf{B} = \omega\mathbf{I}$ and $\mathbf{C}$ is symmetric square ($\mathbf{C} = \mathbf{C}^\top$). By apply these assumption, we get:

Recall the following equality:

$$\begin{aligned}\|\mathbf{z}_k^r\|^2 &= \left\|\left[\begin{pmatrix}\left(\mathbf{I}-\gamma(\omega\mathbf{I})\right) & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{I}-\gamma(\omega\mathbf{I})\right)\end{pmatrix}^k\begin{pmatrix}\mathbf{I} & \omega^{-1}\mathbf{C}\\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I}\end{pmatrix} + \begin{pmatrix}\mathbf{0} & -\omega^{-1}\mathbf{C}\\ \omega^{-1}\mathbf{C}^\top & \mathbf{0}\end{pmatrix}\right]\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\right\|^2\\[2ex] &= \left\|\begin{pmatrix}\left(\mathbf{I}-\gamma(\omega\mathbf{I})\right) & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{I}-\gamma(\omega\mathbf{I})\right)\end{pmatrix}^k\begin{pmatrix}\mathbf{I} & \omega^{-1}\mathbf{C}\\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\right\|^2 + \left\|\begin{pmatrix}\mathbf{0} & -\omega^{-1}\mathbf{C}\\ \omega^{-1}\mathbf{C}^\top & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\right\|^2\\ &\quad + 2\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}^\top\begin{pmatrix}\mathbf{I} & \omega^{-1}\mathbf{C}\\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I}\end{pmatrix}^\top\begin{pmatrix}\left(\mathbf{I}-\gamma(\omega\mathbf{I})\right) & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{I}-\gamma(\omega\mathbf{I})\right)\end{pmatrix}^k\begin{pmatrix}\mathbf{0} & -\omega^{-1}\mathbf{C}\\ \omega^{-1}\mathbf{C}^\top & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\end{aligned}$$

For the last equality we used $\|\mathbf{v}+\mathbf{u}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{u}\|^2 + 2\mathbf{v}^T\mathbf{u}$, where $\mathbf{v}$ and $\mathbf{u}$ are two vectors.

$$\begin{aligned}
\|\mathbf{z}_k^r\|^2 \leqslant{}& \left\| \begin{pmatrix} (\mathbf{I}-\gamma(\omega\mathbf{I})) & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}-\gamma(\omega\mathbf{I})) \end{pmatrix}^k \right\|^2 \left\| \begin{pmatrix} \mathbf{I} & \omega^{-1}\mathbf{C} \\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \mathbf{0} & -\omega_{\mathbf{x}}^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \right\|^2 \\
&+ 2 \left\| \begin{pmatrix} (\mathbf{I}-\gamma(\omega\mathbf{I})) & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}-\gamma(\omega\mathbf{I})) \end{pmatrix}^k \right\| \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
\\
\leqslant{}& (1-\gamma\omega)^{2k} \left\| \begin{pmatrix} \mathbf{I} & \omega^{-1}\mathbf{C} \\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \right\|^2 \\
&+ 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}
\end{aligned}$$

where the second equality follows from the following sequence of reasoning. Using the Lemma 12, we have:

$$\left\| \begin{pmatrix} (\mathbf{I}-\gamma(\omega\mathbf{I})) & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}-\gamma(\omega\mathbf{I})) \end{pmatrix}^k \right\| \leqslant (\max\{1-\gamma\omega, 1-\gamma\omega\})^k \leqslant (1-\gamma\omega)^k$$

$$\begin{aligned}
\|\mathbf{z}_k^r\|^2 \leqslant{}& (1-\gamma\omega)^{2k} \left\| \begin{pmatrix} \mathbf{I} & \omega^{-1}\mathbf{C} \\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \right\|^2 \\
&+ 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
\\
={}& (1-\gamma\omega)^{2k} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & \omega^{-1}\mathbf{C} \\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & \omega^{-1}\mathbf{C} \\ -\omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
&+ \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix}^\top \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
&+ 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
\\
={}& (1-\gamma\omega)^{2k} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I}+\omega^{-2}\mathbf{C}^{\mathbf{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}+\omega^{-2}\mathbf{C}^2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} + \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \omega^{-2}\mathbf{C}^{\mathbf{2}} & \mathbf{0} \\ \mathbf{0} & \omega^{-2}\mathbf{C}^2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
&+ 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \left[\mathbf{I} + \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \right] \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
\\
={}& (1-\gamma\omega)^{2k} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} + \left((1-\gamma)^{2k} + 1 \right) \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \omega^{-2}\mathbf{C}^{\mathbf{2}} & \mathbf{0} \\ \mathbf{0} & \omega^{-2}\mathbf{C}^2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
&+ 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \left[\mathbf{I} + \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \right] \begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
\\
\leqslant{}& (1-\gamma\omega)^{2k} \|\mathbf{z}_0^r\|^2 + \left((1-\gamma\omega)^{2k} + 1 \right) \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \omega^{-2}\mathbf{C}^{\mathbf{2}} & \mathbf{0} \\ \mathbf{0} & \omega^{-2}\mathbf{C}^2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
&+ 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \underbrace{\begin{pmatrix} \mathbf{0} & -\omega^{-1}\mathbf{C} \\ \omega^{-1}\mathbf{C}^\top & \mathbf{0} \end{pmatrix}}_{\phi} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix} \\
&- 2(1-\gamma\omega)^k \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}^\top \begin{pmatrix} \omega^{-2}\mathbf{C}^2 & \mathbf{0} \\ \mathbf{0} & \omega^{-2}\mathbf{C}^2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0^r \\ \mathbf{y}_0^r \end{pmatrix}
\end{aligned}$$

Matrix ϕ is an anti-symmetric matrix and $z^\top \phi z = \mathbf{0}$, assuming ϕ is an anti-symmetric matrix. So we have,

$$
\begin{aligned}
\left\|\mathbf{z}_k^r\right\|^2 &\leqslant (1-\gamma\omega)^{2k}\left\|\mathbf{z}_0^r\right\|^2 + \big((1-\gamma\omega)^{2k} - 2(1-\gamma\omega)^k + 1\big)\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}^\top \begin{pmatrix}\omega^{-2}\mathbf{C}^{\mathbf{2}} & \mathbf{0}\\ \mathbf{0} & \omega^{-2}\mathbf{C}^2\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\\
&= (1-\gamma\omega)^{2k}\left\|\mathbf{z}_0^r\right\|^2 + \big((1-\gamma\omega)^{k} - 1\big)^2\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}^\top \begin{pmatrix}\omega^{-2}\mathbf{C}^{\mathbf{2}} & \mathbf{0}\\ \mathbf{0} & \omega^{-2}\mathbf{C}^2\end{pmatrix}\begin{pmatrix}\mathbf{x}_0^r\\ \mathbf{y}_0^r\end{pmatrix}\\
&\leqslant (1-\gamma\omega)^{2k}\left\|\mathbf{z}_0^r\right\|^2 + \big((1-\gamma\omega)^{k} - 1\big)^2\omega^{-2}\lambda_{\max}^2(\mathbf{C})\left\|\mathbf{z}_0^r\right\|^2\\
&\leqslant \Big((1-\gamma\omega)^{2k} + \big((1-\gamma\omega)^{k} - 1\big)^2\omega^{-2}\lambda_{\max}^2(\mathbf{C})\Big)\left\|\mathbf{z}_0^r\right\|^2
\end{aligned}
$$

Where we used Lemma 12 for the third inequality.

□