Fundamental Benefit of Alternating Updates in Minimax Optimization

Jaewook Lee [*1] **Hanseul Cho** [*1] **Chulhee Yun** [1]

Abstract

The Gradient Descent-Ascent (GDA) algorithm, designed to solve minimax optimization problems, takes the descent and ascent steps either simultaneously (Sim-GDA) or alternately (Alt-GDA). While Alt-GDA is commonly observed to converge faster, the performance gap between the two is not yet well understood theoretically, especially in terms of global convergence rates. To address this theory-practice gap, we present fine-grained convergence analyses of both algorithms for strongly-convex-strongly-concave and Lipschitz-gradient objectives. Our new iteration complexity upper bound of Alt-GDA is strictly smaller than the lower bound of Sim-GDA; *i.e.*, Alt-GDA is provably faster. Moreover, we propose Alternating-Extrapolation GDA (Alex-GDA), a general algorithmic framework that subsumes Sim-GDA and Alt-GDA, for which the main idea is to alternately take gradients from extrapolations of the iterates. We show that Alex-GDA satisfies a smaller iteration complexity bound, identical to that of the Extra-gradient method, while requiring less gradient computations. We also prove that Alex-GDA enjoys linear convergence for bilinear problems, for which both Sim-GDA and Alt-GDA fail to converge at all.

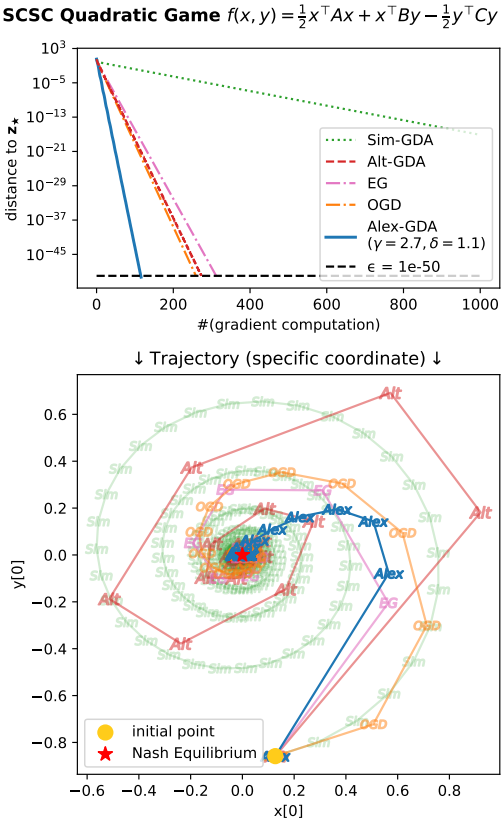

Figure 1. (**Top**) Comparing the convergence speeds of algorithms: **Sim-GDA**, **Alt-GDA**, **EG**, **OGD** and **Alex-GDA**. (**Bottom**) Trajectory of the algorithms. (Partial visualization. Originally, the trajectory is 6-dimensional since $d_x = d_y = 3$).

1. Introduction

The *minimax problem* aims to solve:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d_x}} \max_{\boldsymbol{y}\in\mathbb{R}^{d_y}} f(\boldsymbol{x},\boldsymbol{y}). \tag{1}$$

This has been popularized since the work by von Neumann (1928) and is widely studied in mathematics, economics, computer science, and machine learning. Particularly, in modern machine learning, there are many important settings which fall within problem (1), including but not limited to generative adversarial networks (GANs) (Arjovsky et al., 2017; Goodfellow et al., 2020; Heusel et al., 2017), adversarial training and robust optimization (Latorre et al., 2023; Madry et al., 2018; Sinha et al., 2018; Yu et al., 2022), reinforcement learning (Li et al., 2019), and area-under-curve (AUC) maximization (Liu et al., 2020; Ying et al., 2016; Yuan et al., 2021).

The simplest baseline algorithm for solving minimax problems is *gradient descent-ascent* (GDA) (Dem'yanov & Pevnyi, 1972), which naturally generalizes the idea of gradient descent for minimization problems. The GDA algorithm updates $\boldsymbol{x}$ in the direction of decreasing the objective function f while updating $\boldsymbol{y}$ in the direction of increasing f, either simultaneously (**Sim-GDA**) or alternately (**Alt-GDA**).

[*]Equal contribution [1]KAIST AI, South Korea. Correspondence to: Chulhee Yun <chulhee.yun@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Unfortunately, it is not easy for both algorithms to converge to an optimal point even in a convex-concave minimax problem: in an unconstrained bilinear problem $\min_x \max_y xy$, for example, **Sim-GDA** diverges all the way out while **Alt-GDA** generates bounded but non-convergent iterates (Bailey et al., 2020; Gidel et al., 2019a;b; Zhang et al., 2022).

To tackle the issues of vanilla GDA(s), numerous algorithms have been introduced and analyzed for smooth minimax problems, including Extra-gradient (**EG**) (Korpelevich, 1976), Optimistic Gradient Descent (**OGD**) (Popov, 1980), negative momentum (Gidel et al., 2019b), and many more (Lee & Kim, 2021; Park & Ryu, 2022; Yoon & Ryu, 2021; 2022). Although these algorithms enjoy accelerated convergence rates compared to vanilla GDA, the majority of these works focus on simultaneous updates of $\boldsymbol{x}$ and $\boldsymbol{y}$, mainly because of the simplicity of analysis. However, in minimax problems applied in practical machine learning, it is more natural for the training procedure to work in an alternating sense. In training GANs, for instance, the discriminator should update its weight based on the outcome of the generator, and vice versa. Moreover, there exist substantial amounts of empirical evidence of **Alt-GDA** exhibiting faster convergence (Goodfellow et al., 2020; Mescheder et al., 2017), as we demonstrate in Figure 1. In contrast, we still lack a theoretical understanding of why and how much **Alt-GDA** is faster, especially compared to **Sim-GDA**. To fill this gap between theory and practice, it is a timely and important subject to study which one is a winner between simultaneous and alternating updates.

An existing work by Zhang et al. (2022) proposes a theoretical explanation involving *local* convergence guarantees for μ-strongly-convex-strongly-concave (SCSC), L-Lipschitz gradient functions. Their results constructively explain that **Alt-GDA** (of iteration complexity $\tilde{\mathcal{O}}(\kappa)$) has a faster convergence rate than **Sim-GDA** ($\tilde{\mathcal{O}}(\kappa^2)$), where $\kappa = L/\mu$ is the condition number of the problem. However, their results are confined to guaranteeing *local* convergence rates, which is only valid after enough iterations.

Overall, this raises the following question:

> *For minimax problems* (1)*, are* ***alternating*** *updates strictly better than* ***simultaneous*** *updates, even in terms of* ***global convergence***? (2)

1.1. Summary of Contributions

Our contributions are largely twofold. First, we eliminate the limitations of prior work by providing *global* convergence guarantees that elucidate the fundamental strength of **Alt-GDA** over **Sim-GDA**. Second, we propose a novel algorithm called **Al**ternating-**Ex**trapolation **GDA** (**Alex-GDA**) that achieves an identical rate to the Extra-gradient (**EG**) method with the same number of gradient computations per iteration as **Sim-GDA** and **Alt-GDA**.

For the following results, we assume (μ_x, μ_y)-strongly-convex-strongly-concave (SCSC), (L_x, L_y, L_{xy})-Lipschitz gradient objectives with condition numbers $\kappa_x = L_x/\mu_x$, $\kappa_y = L_y/\mu_y$, and $\kappa_{xy} = L_{xy}/\sqrt{\mu_x\mu_y}$.[1] In particular, we study the upper and lower bounds on the rates of the iteration complexity K required to achieve $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \le \epsilon$.

- In Section 3, we prove that **Sim-GDA** satisfies an iteration complexity rate of

  $$\Theta\left((\kappa_x + \kappa_y + \kappa_{xy}^2)\cdot\log(1/\epsilon)\right)$$

  by showing tightly matching upper and lower bounds. Our fine-grained convergence rate highlights the fact that the term κ_{xy}^2 is the main cause of slow convergence, which previously known results do not capture.

- In Section 4, we prove that **Alt-GDA** satisfies an iteration complexity rate upper bound of

  $$\mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}(\sqrt{\kappa_x} + \sqrt{\kappa_y})\right)\cdot\log(1/\epsilon)\right),$$

  which, compared to the results in Section 3, concludes that ***Alt-GDA** is provably faster than **Sim-GDA***.

- In Section 5, we propose a new algorithm, **Al**ternating-**Ex**trapolation **GDA** (**Alex-GDA**), and prove a smaller iteration complexity rate of

  $$\Theta\left((\kappa_x + \kappa_y + \kappa_{xy})\cdot\log(1/\epsilon)\right)$$

  by showing tightly matching upper and lower bounds. We also show that **EG**—which requires twice the number of gradient computations per iteration—yields the same rate by showing an identical lower bound.

Next, we turn to bilinear objectives $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{y}$, for which both **Sim-GDA** and **Alt-GDA** fail to converge.

- In Section 6, we show that **Alex-GDA** enjoys linear convergence with an iteration complexity rate upper bound of

  $$\mathcal{O}\left((L_{xy}/\mu_{xy})^2\cdot\log(1/\epsilon)\right),$$

  where μ_{xy}, L_{xy} are the smallest, largest nonzero singular values of the coupling matrix $\boldsymbol{B}$, respectively.

Long story short, our results altogether answer the ground-setting question (2) in the positive. For the optimization community—we believe that our fundamental comparison between simultaneous and alternating updates could provide fruitful insights for future investigations to unveil new rate-optimal algorithms by exploiting alternating updates.

[1] For the definitions of SCSC and Lipschitz-gradient functions, please refer to Definitions 2.2 and 2.3. For the definition of condition numbers κ_x, κ_y, and κ_{xy}, please refer to Definition 2.4.

2. Preliminaries

Notation. We study unconstrained minimax problems with objective function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$, where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$ are the variables. In some cases we use $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ for notational simplicity. We denote by $\|\cdot\|$ the Euclidean ℓ_2-norm for vectors and the spectral norm (*i.e.*, maximum singular value) for matrices. We denote by $\langle \cdot, \cdot \rangle$ the usual inner product between vectors in Euclidean space of the same dimension. The spectral radius (*i.e.*, maximum absolute eigenvalue) of a matrix $\boldsymbol{M}$ is denoted by $\rho(\boldsymbol{M})$. The letters $\mathcal{O}$, Ω, ω, and Θ are for the conventional asymptotic notations, while the tilde notation (*e.g.*, $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$) hides polylogarithmic factors.

2.1. Function Class

We first introduce the definitions we need in order to characterize the function class we will mainly focus on.

Definition 2.1 (Strong convexity/concavity)**.** For a given constant $\mu > 0$, we say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is *μ-strongly convex* if

$$f(\boldsymbol{z}') \geq f(\boldsymbol{z}) + \langle \nabla f(\boldsymbol{z}), \boldsymbol{z}' - \boldsymbol{z} \rangle + \frac{\mu}{2}\|\boldsymbol{z}' - \boldsymbol{z}\|^2$$

for all $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$, and *$\mu$-strongly concave* if $-f(\boldsymbol{z})$ is μ-strongly convex. If the above inequality holds for f (or $-f$) and $\mu = 0$, then we say that f is *convex* (or *concave*).

Definition 2.2 (Strong-convex-strong-concavity)**.** For given constants $\mu_x, \mu_y > 0$, we say that a differentiable function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is *(μ_x, μ_y)-strong-convex-strong-concave* (or *(μ_x, μ_y)-SCSC*) if

- $f(\cdot, \boldsymbol{y})$ is μ_x-strongly convex for all $\boldsymbol{y} \in \mathbb{R}^{d_y}$,
- $f(\boldsymbol{x}, \cdot)$ is μ_y-strongly concave for all $\boldsymbol{x} \in \mathbb{R}^{d_x}$.

If $\mu_x = \mu_y = 0$, we say that f is *convex-concave*.

Definition 2.3 (Lipschitz gradients)**.** For given constants $L_x, L_y \geq 0$ and $L_{xy} \geq 0$, we say that a differentiable function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ has *(L_x, L_y, L_{xy})-Lipschitz gradients* if

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}', \boldsymbol{y}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\| \leq L_x \|\boldsymbol{x}' - \boldsymbol{x}\|,$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$,

$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}') - \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})\| \leq L_y \|\boldsymbol{y}' - \boldsymbol{y}\|,$$

for all $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^{d_y}$ and $\boldsymbol{x} \in \mathbb{R}^{d_x}$, and

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}') - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\| \leq L_{xy} \|\boldsymbol{y}' - \boldsymbol{y}\|,$$
$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}', \boldsymbol{y}) - \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})\| \leq L_{xy} \|\boldsymbol{x}' - \boldsymbol{x}\|$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^{d_y}$.

For SCSC and Lipschitz-gradient objective functions, the convergence rates of algorithms usually depend on the ratio between the parameters μ_x, μ_y and L_x, L_y, L_{xy}, which we often refer to as the *condition number*.

Definition 2.4 (Condition numbers)**.** For given constants $0 < \mu_x \leq L_x$, $0 < \mu_y \leq L_y$, and $L_{xy} \geq 0$, we define the *condition numbers* as $\kappa_x := L_x/\mu_x$, $\kappa_y := L_y/\mu_y$, and $\kappa_{xy} := L_{xy}/\sqrt{\mu_x \mu_y}$.

The definitions of κ_x and κ_y are completely analogous to the definition widely used in convex optimization literature, and we have $\kappa_x, \kappa_y \geq 1$ since $\mu_x \leq L_x$, $\mu_y \leq L_y$. The number $\kappa_{xy} \geq 0$ additionally takes into account how the coupling between the two variables can affect the convergence speed.

Definition 2.5 (Function class)**.** For $0 < \mu_x \leq L_x$, $0 < \mu_y \leq L_y$, and $L_{xy} \geq 0$, we define $\mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ as the function class containing all $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ that are (i) *twice-differentiable*, (ii) *(μ_x, μ_y)-SCSC*, and (iii) *has (L_x, L_y, L_{xy})-Lipschitz gradients*.

Considering the minimax problem as in (1), the optimal solution is often characterized as in Definition 2.6.

Definition 2.6. A *Nash equilibrium* of a function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is defined as a point $(\boldsymbol{x}_\star, \boldsymbol{y}_\star) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ which satisfies for all $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$:

$$f(\boldsymbol{x}_\star, \boldsymbol{y}) \leq f(\boldsymbol{x}_\star, \boldsymbol{y}_\star) \leq f(\boldsymbol{x}, \boldsymbol{y}_\star).$$

It is well known that if $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$, then the Nash equilibrium $(\boldsymbol{x}_\star, \boldsymbol{y}_\star)$ of f uniquely exists (see Zhang et al. (2022)).

2.2. Algorithms

We focus on GDA algorithms with constant step sizes $\alpha, \beta > 0$. In Sections 3 and 4, we provide convergence analyses for **Sim-GDA** and **Alt-GDA**, shown in Algorithm 1. In Sections 5 and 6, we construct a new algorithm called **Al**ternating-**Ex**trapolation **GDA** (**Alex-GDA**), shown in Algorithm 2, which we formally define later.

Algorithm 1 **Sim-GDA** and **Alt-GDA**

Input: Number of epochs K, step sizes $\alpha, \beta > 0$
Initialize: $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$
for $k = 0, \ldots, K-1$ **do**
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)$
  if Sim-GDA then
    $\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)$
  else if Alt-GDA then
    $\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k)$
  end if
end for
Output: $(\boldsymbol{x}_K, \boldsymbol{y}_K) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

2.3. Lyapunov Function

Originally designed for stability analysis of dynamical systems (Kalman & Bertram, 1960), the *Lyapunov function* defined as in Definition 2.7 (sometimes referred to as the *potential function*) is widely used as a strategy to obtain convergence guarantees in optimization studies (Bansal & Gupta, 2019; Taylor et al., 2018).

Definition 2.7 (Lyapunov function)**.** Suppose that we have a function f with optimal point $\boldsymbol{z}_\star$, an initialization point $\boldsymbol{z}_0$, and an algorithm that outputs $\boldsymbol{z}_k$ at the k-th iterate. A *Lyapunov function* is defined as a continuous function $\Psi : \mathbb{R}^d \to \mathbb{R}$ such that:

- $\Psi(\boldsymbol{z}) \geq 0$ and $\Psi(\boldsymbol{z}) = 0$ if and only if $\boldsymbol{z} = \boldsymbol{z}_\star$,
- $\Psi(\boldsymbol{z}) \to \infty$ as $\|\boldsymbol{z}\| \to \infty$,
- $\Psi(\boldsymbol{z}_{k+1}) \leq \Psi(\boldsymbol{z}_k)$ for all $k \geq 0$.

For an algorithm that outputs $\{\boldsymbol{z}_k\}_{k\geq 0}$ and a Lyapunov function Ψ, we define the sequence $\{\Psi_k\}_{k\geq 0}$ as $\Psi_k := \Psi(\boldsymbol{z}_k)$, which we will refer to as, with a bit of an abuse of notation, just the *Lyapunov function* throughout the paper.

Definition 2.8. We say that a Lyapunov function $\{\Psi_k\}_{k\geq 0}$ is *valid* if it satisfies $\Psi_k \geq A\|\boldsymbol{z}_k - \boldsymbol{z}_\star\|^2$ for all k and for some constant $A > 0$.

If we find a *valid* Lyapunov function with contraction factor $r \in (0, 1)$— that is, for all $k \geq 0$, we have $\Psi_{k+1} \leq r\Psi_k$, then we can deduce that

$$K = \mathcal{O}\left(\frac{1}{1-r} \cdot \log \frac{\Psi_0}{A\epsilon}\right) \tag{3}$$

iterations are sufficient to ensure $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$. We refer to K as the *iteration complexity*, and the rate in the right-hand side of (3) as the *iteration complexity upper bound*.

3. Convergence Analysis of Sim-GDA

Given an objective function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$, for which the Nash equilibrium is unique, we define the scaled distance to the Nash equilibrium $V(\boldsymbol{x}, \boldsymbol{y})$ as

$$V(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\alpha}\|\boldsymbol{x} - \boldsymbol{x}_\star\|^2 + \frac{1}{\beta}\|\boldsymbol{y} - \boldsymbol{y}_\star\|^2.$$

For **Sim-GDA**, we focus on the convergence rate in terms of the Lyapunov function $\Psi_k^{\text{Sim}} = V(\boldsymbol{x}_k, \boldsymbol{y}_k)$. Note that Ψ_k^{Sim} is always nonnegative, and is valid since we have $A^{\text{Sim}}\|\boldsymbol{z}_k - \boldsymbol{z}_\star\|^2 \leq \Psi_k^{\text{Sim}}$ for $A^{\text{Sim}} = \min\left\{\frac{1}{\alpha}, \frac{1}{\beta}\right\}$. This potential function is a popular choice in minimax optimization or variance reduction problems with step sizes of different scales (Palaniappan & Bach, 2016).

3.1. Convergence Upper Bound

Theorem 3.1 yields a contraction result for **Sim-GDA**.

Theorem 3.1. *Suppose that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$. Then, there exists a pair of step sizes α, β with*

$$\alpha\mu_x = \beta\mu_y = \Theta\left(\frac{1}{\kappa_x + \kappa_y + \kappa_{xy}^2}\right),$$

*such that **Sim-GDA** satisfies $\Psi_{k+1}^{\text{Sim}} \leq r\Psi_k^{\text{Sim}}$ with*

$$r = \left(\frac{\left(\kappa_{xy} + \sqrt{\max\{\kappa_x, \kappa_y\} + \kappa_{xy}^2}\right)^2 - 1}{\left(\kappa_{xy} + \sqrt{\max\{\kappa_x, \kappa_y\} + \kappa_{xy}^2}\right)^2 + 1}\right)^2. \tag{4}$$

While we defer the proof of Theorem 3.1 to Appendix B.1, by (3) we can restate the convergence rate upper bound in terms of the iteration complexity as follows.

Corollary 3.2. *For the step sizes given as in Theorem 3.1, **Sim-GDA** linearly converges with iteration complexity*

$$\mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right) \cdot \log \frac{\Psi_0^{\text{Sim}}}{A^{\text{Sim}}\epsilon}\right),$$

where $A^{\text{Sim}} = \min\left\{\frac{1}{\alpha}, \frac{1}{\beta}\right\}$.

We defer the proof of Corollary 3.2 to Appendix B.2.

Comparison with Previous Work. The previously known iteration complexity upper bound of **Sim-GDA** was $\tilde{\mathcal{O}}(\kappa^2)$ (Mescheder et al., 2017; Azizian et al., 2020; Zhang et al., 2022), where the condition number is defined as $\kappa = \frac{\max\{L_x, L_y, L_{xy}\}}{\min\{\mu_x, \mu_y\}}$. However, using a single condition number might oversimplify the problem and lead to loose results; for instance, if the condition numbers follow $\kappa_x, \kappa_y = \Theta(t^2)$ and $\kappa_{xy} = \Theta(t)$ for some t, then previous results can only guarantee up to $\tilde{\mathcal{O}}(t^4)$, while Corollary 3.2 suggests a better rate of $\tilde{\mathcal{O}}(t^2)$. This shows that separating the condition numbers helps capture how κ_{xy}, or the *interaction between $\boldsymbol{x}$ and $\boldsymbol{y}$*, affects convergence speed.

Meanwhile, a recent work by Zamani et al. (2022) proposes an iteration complexity upper bound for **Sim-GDA** of $\tilde{\mathcal{O}}(\overline{\kappa} + \kappa_{xy}^2)$ for $\overline{\kappa} = \frac{\max\{L_x, L_y\}}{\min\{\mu_x, \mu_y\}}$, but the proof heavily relies on a computer-assisted method known as the Performance Estimation Problem (PEP) (Drori & Teboulle, 2014). Our fine-grained analysis subsumes all of these previous results, and—to the best of our knowledge—is the first to clarify the exact convergence rate of **Sim-GDA** in terms of individual condition numbers κ_x, κ_y, and κ_{xy}.

3.2. Convergence Lower Bound

Theorem 3.3 provides a convergence lower bound of the iteration complexity of **Sim-GDA** which holds for all possible step sizes $\alpha, \beta > 0$.

Theorem 3.3. *There exists a 6-dimensional function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ with $d_x = d_y = 3$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of **Sim-GDA** requires an iteration complexity of rate at least*

$$\Omega\left(\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$.

The iteration complexity lower bound in Theorem 3.3 exactly matches the upper bound in Corollary 3.2, ensuring that our analysis on **Sim-GDA** is tight (ignoring log factors). We defer the proof of Theorem 3.3 to Appendix B.3.

Remark. Unlike typical lower bounds for which the initialization is specifically chosen along with the function, our results in Theorem 3.3 works for any initialization, while the dependency on initialization is hidden in the numerator in the $\log(1/\epsilon)$ part similarly as in the upper bound results. All we need is an initialization point with $\mathcal{O}(1)$ distance from the optimum, and the same applies to the lower bound results we present in Theorem 5.3.

4. Convergence Analysis of Alt-GDA

For **Alt-GDA**, the half-step iterates alternating between $\boldsymbol{x}$ and $\boldsymbol{y}$ updates make theoretical analysis much harder than when dealing with simultaneous updates. We address this by focusing on the convergence rate in terms of the following Lyapunov function (instead of just $V(\boldsymbol{x}_k, \boldsymbol{y}_k)$):

$$\begin{aligned}\Psi_k^{\text{Alt}} = V^{\text{Alt}}(\boldsymbol{x}_k, \boldsymbol{y}_k) + V^{\text{Alt}}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k) \\ - \alpha(1 - \alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2,\end{aligned}$$

where $V^{\text{Alt}}(\boldsymbol{x}, \boldsymbol{y})$ is defined as

$$\left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x} - \boldsymbol{x}_\star\|^2 + \left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y} - \boldsymbol{y}_\star\|^2.$$

Note that we capture the two-step-alternating nature of the algorithm by considering two adjacent iterates at a time, which turns out to be the key idea in the proofs.

4.1. Convergence Upper Bound

Theorem 4.1 yields a contraction result for **Alt-GDA**.

Theorem 4.1. *Suppose that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and we run **Alt-GDA** with step sizes $\alpha, \beta > 0$ that satisfy*

$$\alpha \leq \frac{1}{2} \cdot \min\left\{\frac{1}{L_x}, \frac{\sqrt{\mu_y}}{L_{xy}\sqrt{L_x}}\right\},$$
$$\beta \leq \frac{1}{2} \cdot \min\left\{\frac{1}{L_y}, \frac{\sqrt{\mu_x}}{L_{xy}\sqrt{L_y}}\right\}.$$

Then Ψ_k^{Alt} is valid, and satisfies $\Psi_{k+1}^{\text{Alt}} \leq r\Psi_k^{\text{Alt}}$ with

$$r = \max\left\{\frac{\frac{1}{\alpha} - \mu_x}{\frac{1}{\alpha} - 2\beta^2 L_y L_{xy}^2}, \frac{\frac{1}{\beta} - \mu_y}{\frac{1}{\beta} - \alpha^2 L_x L_{xy}^2}, \frac{\frac{1}{\alpha} - \mu_x}{\frac{1}{\alpha}}\right\},$$

where we have $0 < r < 1$.

While we defer the proof of Theorem 4.1 to Appendix C.1, by (3) we can restate the convergence rate upper bound in terms of the iteration complexity as follows.

Corollary 4.2. *For step sizes given by the maximum possible values in Theorem 4.1, **Alt-GDA** linearly converges with iteration complexity*

$$\mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}(\sqrt{\kappa_x} + \sqrt{\kappa_y})\right) \cdot \log \frac{\Psi_0^{\text{Alt}}}{A^{\text{Alt}}\epsilon}\right),$$

where $A^{\text{Alt}} = \min\left\{\frac{1}{2\alpha} - \mu_x, 2\left(\frac{3}{4\beta} - \mu_y\right)\right\} > 0$.

We defer the proof of Corollary 4.2 to Appendix C.2.

Recall that for **Sim-GDA** we have an upper bound of $\tilde{\mathcal{O}}\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right)$, and a lower bound which shows that this rate cannot be improved. Comparing this with Corollary 4.2, we can conclude that the convergence rate of **Alt-GDA** is faster than **Sim-GDA**.

Comparison with Sim-GDA. Our fine-grained analysis clarifies how the dependence of the convergence speed of **Sim-GDA** and **Alt-GDA** on κ_{xy}, corresponding to the *interaction between $\boldsymbol{x}$ and $\boldsymbol{y}$*, are different from each other. If $\kappa_{xy} = O(\sqrt{\kappa_x} + \sqrt{\kappa_y})$, then the diagonal blocks of the Hessian dominate, for which both **Sim-GDA** and **Alt-GDA** exhibit similar convergence dynamics to plain GD. If $\kappa_{xy} = \omega(\sqrt{\kappa_x} + \sqrt{\kappa_y})$, i.e., the off-diagonal block dominates, then the relatively large interaction between $\boldsymbol{x}$ and $\boldsymbol{y}$ slows down convergence. Our results show that **Alt-GDA** is capable of faster convergence essentially because its dependency on κ_{xy} is of smaller order.

Comparison with Local Analysis. Zhang et al. (2022) show that the *local* convergence rates of **Sim-GDA** and **Alt-GDA** are $\tilde{\mathcal{O}}(\kappa^2)$ and $\tilde{\mathcal{O}}(\kappa)$, respectively, where $\kappa = \frac{\max\{L_x, L_y, L_{xy}\}}{\min\{\mu_x, \mu_y\}}$. Such kinds of *local* convergence rates of operators, including GDA iterates, rely on (the spectral radius of) the Jacobian matrix of the operator at the optimum (Bertsekas, 1999) and require that the iterates are in a small neighborhood around the optimum, or—for gradient methods—that the objective function is *quadratic*, so that the Jacobian is constant and the same spectral arguments hold everywhere in the domain. In contrast, Corollaries 3.2 and 4.2 both show *global* convergence rates for all initialization and SCSC objectives without such assumptions.

While we can see that Corollary 3.2 naturally subsumes the local convergence rate $\tilde{\mathcal{O}}(\kappa^2)$, it turns out that Corollary 4.2 is analogous to $\tilde{\mathcal{O}}(\kappa^{3/2})$, which is has a gap of $\kappa^{1/2}$ with the local convergence rate of $\tilde{\mathcal{O}}(\kappa)$ by Zhang et al. (2022). Viewing the local convergence result as a global convergence bound for the smaller class of *quadratic* SCSC functions, we believe that there may exist a *non-quadratic* function for which **Alt-GDA** requires an iteration complexity of $\tilde{\omega}(\kappa)$, which we discuss in detail in Conjecture 8.1.

5. Alternating-Extrapolation GDA

A natural way of unifying the baseline algorithms **Sim-GDA** and **Alt-GDA** is to think of taking a *linear combination* between the two. That is, we can write:

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k),\\ \tilde{\boldsymbol{x}}_{k+1} &= (1-\gamma)\boldsymbol{x}_k + \gamma\boldsymbol{x}_{k+1},\\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{y}_k).\end{aligned} \tag{5}$$

Note that this formulation provides an *interpolation* between **Sim-GDA** ($\gamma = 0$) and **Alt-GDA** ($\gamma = 1$). In the previous sections, we demonstrated a provable gap in the iteration complexity between the two endpoints $\gamma = 0$ and 1; this motivates us to consider an *extrapolation* to $\gamma > 1$ and see if we can achieve a further speed-up.

However, if we extrapolate the $\boldsymbol{x}$ side alone, the update equations for $\boldsymbol{x}$ and $\boldsymbol{y}$ will no longer be of the same form. By symmetrizing the $\boldsymbol{x}$ and $\boldsymbol{y}$ sides, we now obtain the following general framework:

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k),\\ \tilde{\boldsymbol{x}}_{k+1} &= (1-\gamma)\boldsymbol{x}_k + \gamma\boldsymbol{x}_{k+1},\\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{y}_k),\\ \tilde{\boldsymbol{y}}_{k+1} &= (1-\delta)\boldsymbol{y}_k + \delta\boldsymbol{y}_{k+1},\end{aligned}$$

where $\tilde{\boldsymbol{x}}_{k+1}$ and $\tilde{\boldsymbol{y}}_{k+1}$ are the points where we compute the gradients, and $\gamma, \delta \geq 0$ are hyperparameters. Notice that choosing $(\gamma, \delta) = (0, 1)$ recovers **Sim-GDA** and $(\gamma, \delta) = (1, 1)$ corresponds to **Alt-GDA**.

We can rewrite our updates in terms of gradient updates (Algorithm 2). We name our algorithm framework **Al**ternating-**Ex**trapolation **GDA** (**Alex-GDA**), after the fact that our analysis mainly focuses on the case $\gamma, \delta > 1$ in which we compute gradients using *extrapolated* iterates, and we make alternating updates between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Initialization. Some careful readers might notice that the first step of **Alex-GDA** is a bit different from the rest of the iterations; for $k = 0$ we set $\tilde{\boldsymbol{y}}_0 = \boldsymbol{y}_0$, whereas we use $\tilde{\boldsymbol{y}}_k = \boldsymbol{y}_k + (\delta - 1)\beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})$ for all subsequent steps ($k \geq 1$). This requires a bit more careful analysis, just as in how we define the Lyapunov function for **Alex-GDA**:

$$\begin{aligned}\Psi_k^{\text{Alex}} &= V(\boldsymbol{x}_k, \boldsymbol{y}_k) + V(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k)\\ &- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)\|^2 + (\delta - 1)\beta\|\nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2\\ &+ \frac{(\gamma - 1)(\delta - 1)\alpha\beta}{1 - \alpha\mu_x} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1}, \tilde{\boldsymbol{y}}_{k-1})\|^2\end{aligned}$$

for $k \geq 1$, and

$$\begin{aligned}\Psi_0^{\text{Alex}} &= V(\boldsymbol{x}_0, \boldsymbol{y}_0) + V(\boldsymbol{x}_1, \boldsymbol{y}_0) - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2\\ &+ \frac{(\gamma - 1)(\delta - 1)\alpha\beta}{(1 - \alpha\mu_x)(1 - \beta\mu_y)} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2\end{aligned}$$

for $k = 0$.

Algorithm 2 **Al**ternating-**Ex**trapolation **GDA** (**Alex-GDA**)

Input: Number of epochs K, step sizes $\alpha, \beta > 0$, hyperparameters $\gamma, \delta \geq 0$
Initialize: $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ and $\tilde{\boldsymbol{y}}_0 = \boldsymbol{y}_0 \in \mathbb{R}^{d_y}$
for $k = 0, \ldots, K - 1$ **do**
$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)$
$\tilde{\boldsymbol{x}}_{k+1} = \boldsymbol{x}_k - \gamma\alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)$
$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{y}_k)$
$\tilde{\boldsymbol{y}}_{k+1} = \boldsymbol{y}_k + \delta\beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{y}_k)$
end for
Output: $(\boldsymbol{x}_K, \boldsymbol{y}_K) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

5.1. Convergence Upper Bound

Theorem 5.1 yields a contraction result for **Alex-GDA**.

Theorem 5.1. *Suppose that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and we run **Alex-GDA** with $\gamma, \delta > 1$ and step sizes $\alpha, \beta > 0$ that satisfy*

$$\alpha \leq C \cdot \min\left\{\frac{1}{L_x}, \frac{\sqrt{\mu_y}}{L_{xy}\sqrt{\mu_x}}\right\},$$
$$\beta \leq C \cdot \min\left\{\frac{1}{L_y}, \frac{\sqrt{\mu_x}}{L_{xy}\sqrt{\mu_y}}\right\}.$$

for some constant $C > 0$ (which only depends on γ and δ). Then Ψ_k^{Alex} is valid, and satisfies $\Psi_{k+1}^{\text{Alex}} \leq r\Psi_k^{\text{Alex}}$ with

$$r = \max\{1 - \alpha\mu_x, 1 - \beta\mu_y\}.$$

While we defer the proof of Theorem 5.1 to Appendix D.1, by (3) we can restate the convergence rate upper bound in terms of the iteration complexity as follows.

Corollary 5.2. *For step sizes given by the maximum possible values in Theorem 5.1, **Alex-GDA** linearly converges with iteration complexity*

$$\mathcal{O}\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log\frac{\Psi_0^{\text{Alex}}}{A^{\text{Alex}}\epsilon}\right),$$

where $A^{\text{Alex}} = \min\left\{\frac{1}{2\alpha}, \frac{1}{\beta}\right\} > 0$.

While we defer the proof of Corollary 5.2 to Appendix D.2, we can observe that Corollary 5.2 provides a stronger iteration complexity upper bound than Corollary 4.2.

5.2. Convergence Lower Bound

Theorem 5.3 provides a convergence lower bound of the iteration complexity of **Alex-GDA** which holds for all possible step sizes $\alpha, \beta > 0$.

Theorem 5.3. *There exists a 6-dimensional function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ with $d_x = d_y = 3$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of **Alex-GDA** with $\gamma, \delta > 1$ requires an iteration complexity of*

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$.

The iteration complexity rate in Theorem 5.3 exactly matches the upper bound in Corollary 5.2, which ensures that our analysis on **Alex-GDA** is tight (ignoring log factors). We defer the proof of Theorem 5.3 to Appendix D.3.

5.3. Comparison with EG

Here we compare **Alex-GDA** to the Extra-gradient (**EG**) method (Korpelevich, 1976), an algorithm based on *simultaneous* updates of the form:

$$\left.\begin{aligned} \boldsymbol{x}_{k+\frac{1}{2}} &= \boldsymbol{x}_k - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \\ \boldsymbol{y}_{k+\frac{1}{2}} &= \boldsymbol{y}_k + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \end{aligned}\right\} \text{exploration steps}$$

$$\left.\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}), \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}). \end{aligned}\right\} \text{update steps}$$

It is known by Mokhtari et al. (2019) that **EG** converges with iteration complexity $\tilde{\mathcal{O}}(\kappa)$, where $\kappa = \frac{\max\{L_x, L_y, L_{xy}\}}{\min\{\mu_x, \mu_y\}}$. While **EG** is famous for its simplicity and fast convergence, we can show that **EG** must satisfy the same lower bound with **Alex-GDA** via the following proposition.

Proposition 5.4. *There exists a 6-dimensional function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ with $d_x = d_y = 3$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of **EG** requires an iteration complexity of rate at least*

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$.

We defer the proof of Proposition 5.4 to Appendix D.4.

By comparing Proposition 5.4 with the upper (and lower) bound for **Alex-GDA**, it is clear that **EG** *cannot* be strictly faster than **Alex-GDA** in terms of iteration complexity rates. Moreover, **Alex-GDA** requires only two gradient values (one for $\boldsymbol{x}$, $\boldsymbol{y}$ each) per a single iteration, while **EG** needs to perform exactly twice the amount of computations (two for $\boldsymbol{x}$, $\boldsymbol{y}$ each). Nevertheless, **Alex-GDA** is provably as fast as **EG**, and in fact, it showcases faster *empirical* convergence compared to **EG** as shown in Figure 1.

In Appendix A, we also compare **Alex-GDA** with another well-known baseline algorithm, Optimistic Gradient Descent (**OGD**) (Popov, 1980).

6. Alex-GDA Converges on Bilinear Problems

One drawback shared by **Sim-GDA** and **Alt-GDA** is that both algorithms fail to converge for simple unconstrained bilinear problems of the form $\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{y}$ (Gidel et al., 2019b), an important special case of a convex-concave but non-SCSC problem with Lipschitz gradients.

Surprisingly, we show that **Alex-GDA**, on the other hand, *does* converge on bilinear problems. In order to present the result, we define μ_{xy} as the smallest *nonzero* singular value of $\boldsymbol{B}$. Note that it is natural to assume the existence of nonzero singular values—if not, then $\boldsymbol{B} = \boldsymbol{0}$, and the objective is constantly zero. Analogously to previous definitions, we choose L_{xy} as the largest singular value of $\boldsymbol{B}$.

We first characterize the exact condition for convergent step sizes of **Alex-GDA** on bilinear problems. Interestingly, it allows a larger range of parameters γ and δ: we no longer require $\gamma > 1$ and $\delta > 1$ here.

Theorem 6.1. *With a proper choice of step sizes α and β, **Alex-GDA** linearly converges to a Nash equilibrium of a bilinear problem if and only if $\gamma + \delta > 2$. In this case, the exact conditions for convergent step sizes α and β are:*

$$\begin{cases} \alpha\beta < \frac{4}{(2\gamma-1)(2\delta-1)L_{xy}^2}, & \text{if } 4\gamma\delta - 3(\gamma+\delta) + 2 \geq 0, \\ \alpha\beta < \frac{\gamma+\delta-2}{-(\gamma-1)(\delta-1)(\gamma+\delta-1)L_{xy}^2}, & \text{if } 4\gamma\delta - 3(\gamma+\delta) + 2 < 0. \end{cases}$$

We defer the proof of Theorem 6.1 to Appendix E.1.

Furthermore, if we properly choose the step size, we can obtain the iteration complexity of **Alex-GDA** on bilinear problems.

Theorem 6.2. *For general $\gamma \geq 1$ and $\delta \geq 1$ such that $\gamma + \delta > 2$, If we choose the step sizes α and β so that $\alpha\beta = \frac{1}{C_{\gamma,\delta} L_{xy}^2}$ where $C_{\gamma,\delta} > 0$ is a constant that only depends on γ and δ, an iteration complexity upper bound of **Alex-GDA** is*

$$\mathcal{O}\left(\frac{C_{\gamma,\delta}}{\gamma + \delta - 2} \cdot \left(\frac{L_{xy}}{\mu_{xy}}\right)^2 \cdot \log\left(\frac{\|\boldsymbol{w}_0\|^2}{\epsilon}\right)\right),$$

where $\|\boldsymbol{w}_0\|^2 = \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2$ and $\boldsymbol{z}_\star = (\boldsymbol{x}_\star, \boldsymbol{y}_\star)$ is a uniquely determined Nash equilibrium if $\boldsymbol{z}_0$ is given.

If $\delta = 1$, the optimal rate exponent of Alex-GDA is

$$\lim_{k\to\infty} \frac{\|\boldsymbol{z}_k - \boldsymbol{z}_\star\|}{\|\boldsymbol{z}_{k-1} - \boldsymbol{z}_\star\|} = \sqrt{\frac{L_{xy}^2 - \mu_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}},$$

where the optimal choice of parameters are

$$\alpha\beta = \frac{2\mu_{xy}^2/L_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}, \quad \gamma = 1 + \frac{L_{xy}^2}{\mu_{xy}^2}.$$

While we defer the proof of Theorem 6.2 to Appendix E.2, we remark that the convergence speed depends on a new type of condition number, namely $L_{xy}/\mu_{xy} \geq 1$, which is distinct from our definition of κ_{xy}.

6.1. Comparison with EG

A work by Zhang & Yu (2020) analyzes optimal convergence rates of **EG** and several other minimax optimization algorithms on bilinear problems. They prove that the optimal rate exponent of **EG** is $\frac{L_{xy}^2 - \mu_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}$, which boils down to the iteration complexity $\tilde{\mathcal{O}}\left((L_{xy}/\mu_{xy})^2\right)$; it matches the iteration complexity of **Alex-GDA** up to constant factor.

It seems that the optimal rate exponent of **EG** is quadratically better than that of **Alex-GDA** with $\delta = 1$. However, since **EG** takes *twice* more gradient computation per iteration than **Alex-GDA**, the optimal *gradient computation complexity* of **EG** and **Alex-GDA** with $\delta = 1$ are exactly identical. Still, there is room for further improvement in the convergence rate of **Alex-GDA** by choosing δ other than 1, but we leave it as a future work.

We also compare **Alex-GDA** with **OGD** in Appendix A.

7. Experiments

The details of the experiments are illustrated in Appendix G.

7.1. SCSC Quadratic Games

An SCSC quadratic game is a minimax problem:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y},$$

where $\boldsymbol{A}$ and $\boldsymbol{C}$ are positive definite matrices.

(1) Small-scale. We conducted experiments on a $(3+3)$-dimensional SCSC quadratic game to visually compare the convergence speed of the algorithms in Figure 1. We choose appropriate step sizes for each algorithm by applying grid search, regarding the number of gradient computations to arrive at an ϵ-distant point from the Nash equilibrium, among convergent step sizes. As shown in the figure and already observed in Zhang et al. (2022), **Alt-GDA** beats **Sim-GDA** in terms of the convergence rate. We additionally observe that the gradient complexity of **Alt-GDA** seems comparable to that of **EG** and **OGD**.[2] Furthermore, with moderately tuned parameters γ and δ, our **Alex-GDA** achieves a convergence rate that is even faster than **EG** and **OGD**.

(2) Higher Dimension, Extensive Comparisons. We run further experiments on $(100+100)$-dimensional SCSC quadratic games to extensively compare GDA, **EG**, **OGD**, and **Alex-GDA**. We test both simultaneous/alternating versions and (either positive or negative) momentum variants. In particular, we investigate five different configurations of problem parameters $(\mu_x = \mu_y, \mu_{xy}, L_x = L_y, L_{xy})$, where μ_{xy} is the smallest singular value of the matrix $\boldsymbol{B}$. The results are shown in Table 1. We observe **Alt-GDA** is much faster than **Sim-GDA** and even faster than **Sim-GDA** with momentum. Among algorithms *without* momentum, **Alex-GDA** exhibits the best gradient complexity. If we include algorithms with momentum, a variant of **Alex-GDA** (Algorithm 3 in Appendix G.2) achieves the best performance among all compared algorithms, while the alternating & momentum variant of **OGD** showcases the second-best performance for most of problem parameters. Lastly, we verify our theoretical findings by observing an increasing trend of gradient complexity in terms of condition numbers $L/\mu (= \kappa_x = \kappa_y)$ and $L_{xy}/\sqrt{\mu_x\mu_y} (= \kappa_{xy})$, but not in terms of L_{xy}/μ_{xy} (introduced for analysis of bilinear problems).

7.2. Generative Adversarial Networks

To examine the efficacy of **Alex-GDA**, we train WGAN-GP (Arjovsky et al., 2017; Gulrajani et al., 2017) for the image generation task, mostly following the implementation details in Heusel et al. (2017). We examine the natural combinations of Adam (Kingma & Ba, 2015) and (stochastic variants of) **Sim-**/**Alt-**/**Alex-GDA**, which we call Sim-/Alt-/Alex-Adam, respectively. We highlight that **Alex-GDA** can be easily implemented on top of any existing base optimizers including Adam because all we need to implement additionally is a couple of extrapolation steps; we provide a brief PyTorch (Paszke et al., 2019) implementation of Alex-Adam for GANs in Listing 1 of Appendix G.3. We moderately tune the step sizes and the values of γ and δ. As a result, we use $(\gamma, \delta) = (1, 4)$ for MNIST (Deng, 2012), $(\gamma, \delta) = (1, 1.2)$ for CIFAR-10 (Krizhevsky et al., 2009), and $(\gamma, \delta) = (1, 2)$ for LSUN Bedroom 64×64 dataset (Yu et al., 2015). The result is shown in Table 2, where we report Fréchet inception distance (FID) scores (Heusel et al., 2017). To the best of our knowledge, we achieve

[2] In all experiments, we allow **EG** with different step sizes used at the exploration and update steps, which is of a more general formulation than the description in Section 5.3. We also use a more general formulation of **OGD** than that explained in Appendix A. See Appendix G.1.

Table 1. **(100+100)-dimensional SCSC quadratic games.** We report the number of gradient computations, averaged over 30 runs. Every algorithm was run until the squared distance from the optimum reached $\leq \epsilon$. We set $\mu_x = \mu_y = \mu$, $L_x = L_y = L$. Note that **Sim** means **Sim-GDA** and **Alt** means **Alt-GDA**. Also, **+M** means momentum (positive or negative), while **+A** means alternating updates. For each row, we mark the first, second, and third places as $*$, †, and ‡, respectively.

$(\mu, \mu_{xy}, L, L_{xy}, \epsilon)$	Sim	Sim +M	Alt	Alt +M	EG	EG +M	EG +A	EG +AM	OGD	OGD +M	OGD +A	OGD +AM	Alex	Alex +M
$(0.1, 0.1, 1, 1, 10^{-8})$	1974.2	421.0	105.9	78.9	133.8	115.6	139.9	102.0	132.8	105.3	90.0	67.6‡	62.7†	44.7*
$(0.1, 0.05, 1, 2, 10^{-8})$	7865.0	839.8	149.1	105.6	253.2	210.6	278.9	186.9	215.1	177.3	116.1	93.6†	100.6‡	69.1*
$(0.01, 0.001, 1, 0.5, 10^{-4})$	42762.1	3824.3	394.9	182.0	291.1	225.9	380.7	228.4	281.1	176.1	182.3	127.7†	133.1‡	58.3*
$(0.01, 0.01, 1, 1, 10^{-4})$	104220.5	8539.4	567.6	157.2	308.8	223.9	299.9	175.7	280.5	184.9	200.5	117.3†	138.8‡	73.2*
$(0.01, 0.05, 1, 2, 10^{-4})$	416822.5	16719.4	777.4	149.0	347.5	253.9	363.6	231.0	337.6	213.3	162.0	108.6†	135.4‡	83.9*

Table 2. **WGAN-GP.** We report the mean (and standard deviation) of FID scores (the lower the better).

	MNIST	CIFAR-10	LSUN Bedroom
Sim-Adam	3.97 (1.3)	45.0 (1.2)	131.2 (8.4)
Alt-Adam	1.85 (0.3)	24.2 (1.7)	9.0 (1.2)
Alex-Adam	**1.53 (0.3)**	**23.8 (1.5)**	**6.3 (0.6)**

state-of-the-art image generation performance in terms of FID scores for MNIST and LSUN Bedroom 64×64 datasets with Alex-Adam.

The experiments in Tables 1 and 2 can be reproduced with our code available at GitHub.[3]

8. Conclusion

We present global convergence rates of **Sim-GDA** and **Alt-GDA** on SCSC, Lipschitz-gradient objectives in terms of the condition numbers κ_x, κ_y, and κ_{xy}. For **Sim-GDA** we prove an iteration complexity of $\tilde{\Theta}(\kappa_x + \kappa_y + \kappa_{xy}^2)$, while for **Alt-GDA** we obtain a smaller iteration complexity of $\tilde{\mathcal{O}}(\kappa_x + \kappa_y + \kappa_{xy}(\sqrt{\kappa_x} + \sqrt{\kappa_y}))$. Comparing the results, we show that **Alt-GDA** is provably faster than **Sim-GDA** in terms of global convergence.

Moreover, we propose a novel algorithm called **Alex-GDA**, inspired by an extension of **Sim-GDA** and **Alt-GDA** via linear extrapolation. **Alex-GDA** shows a faster iteration complexity of $\tilde{\Theta}(\kappa_x + \kappa_y + \kappa_{xy})$, matching the convergence rate of **EG** with less gradient computations per iteration. We also show that **Alex-GDA** converges linearly for bilinear problems, for which **Sim-GDA** and **Alt-GDA** diverge.

We believe that our results, altogether, are valuable demonstrations of ***the benefit of alternating updates*** in GDA algorithms for minimax optimization.

Future Work. As an effort to check if it is possible to obtain $\mathcal{O}(\kappa)$ convergence of **Alt-GDA**, we have tried using a computer-assisted method called the performance estimation problem (PEP) (Drori & Teboulle, 2014), a powerful tool originally designed to infer *tight* worst-case complexities of *convex* optimization algorithms. Based on the work by Das Gupta et al. (2023), we devised a PEP-based tool that automatically finds the worst-case convergence rate of an algorithm by optimizing the function, step size, and performance measure altogether. While it is known by Ryu et al. (2020) that the extension of such methods to *minimax* optimization can only yield a possibly loose *upper bound*, the estimate we obtained for **Alt-GDA** was approximately $\mathcal{O}(\kappa^{1.4})$. Moreover, the estimated rate for **Sim-GDA** was $\mathcal{O}(\kappa^{1.99})$, which is very close to our theoretical results. You may refer to Appendix H for more details.

Based on these observations and the discussions about Theorem 4.1 at the end of Section 4, we leave the following conjecture on the convergence lower bound of **Alt-GDA**.

Conjecture 8.1. *There exists a* ***non-quadratic*** *function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of* ***Alt-GDA*** *requires an iteration complexity of*

$$\Theta\left((\kappa_x + \kappa_y + \kappa_{xy}(\kappa_x + \kappa_y)^p) \cdot \log \frac{1}{\epsilon}\right)$$

for $p \in (0, \frac{1}{2})$.

Also, on top of our findings on bilinear functions in Section 6, we also leave the following conjecture on **Alex-GDA** on general *convex-concave* objectives for future work.

Conjecture 8.2. *Suppose that the objective function f is convex-concave and has (L_x, L_y, L_{xy})-Lipschitz gradients. Then, we conjecture that* ***Alex-GDA*** *exhibits last-iterate convergence to a Nash equilibrium of f.*

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & evaluation (IITP) grant (No. RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)) funded by the Korea government (MSIT). The work was also supported by the National Research Foundation of Korea (NRF) grant (No. RS-2023-00211352) funded by the Korea government (MSIT). CY acknowledges support from a grant funded by Samsung Electronics Co., Ltd.

[3] github.com/HanseulJo/Alex-GDA

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pp. 214–223. PMLR, 2017. 1, 7.2

Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. Accelerating smooth games by manipulating spectral shapes. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1705–1715. PMLR, 2020. URL http://proceedings.mlr.press/v108/azizian20a.html. 3.1, G.2

Bailey, J. P., Gidel, G., and Piliouras, G. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Conference on Learning Theory (COLT)*, pp. 391–407. PMLR, 2020. 1

Bansal, N. and Gupta, A. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019. 2.3

Bertsekas, D. *Nonlinear Programming*. Athena Scientific, 1999. 4.1

Das Gupta, S., Van Parys, B. P. G., and Ryu, E. K. Branch-and-bound performance estimation programming: a unified methodology for constructing optimal optimization methods. *Math. Program.*, 204(1–2): 567–639, jun 2023. ISSN 0025-5610. doi: 10.1007/s10107-023-01973-1. URL https://doi.org/10.1007/s10107-023-01973-1. 8, H

Dem'yanov, V. and Pevnyi, A. Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52, 1972. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(72)90002-X. URL https://www.sciencedirect.com/science/article/pii/004155537290002X. 1

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 7.2, G.3

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1-2):451–482, 2014. doi: 10.1007/s10107-013-0653-0. URL https://doi.org/10.1007/s10107-013-0653-0. 3.1, 8, H

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019a. URL https://openreview.net/forum?id=r1laEnA5Ym. 1

Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1802–1811. PMLR, 2019b. 1, 6, G.2

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

Grove, E. A. and Ladas, G. *Periodicities in nonlinear difference equations*, volume 4. CRC Press, 2004. D.3, F.1

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 7.2

Haynsworth, E. V. On the schur complement. *Basel Mathematical Notes*, 20:17, 1968. E.1

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1, 7.2, G.3

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, Cambridge, England, 2 edition, October 2012. B.4

Kalman, R. E. and Bertram, J. E. Control System Analysis and Design Via the "Second Method" of Lyapunov: I—Continuous-Time Systems. *Journal of Basic Engineering*, 82(2):371–393, 06 1960. ISSN 0021-9223. doi: 10.1115/1.3662604. URL https://doi.org/10.1115/1.3662604. 2.3

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 7.2, G.3

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. 1, 5.3

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images, 2009. 7.2, G.3

Latorre, F., Krawczuk, I., Dadi, L. T., Pethick, T., and Cevher, V. Finding actual descent directions for adversarial training. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=I3HCE7Ro78H. 1

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:22588–22600, 2021. 1

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., and Russell, S. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Conference on Artificial Intelligence (AAAI)*, volume 33, pp. 4213–4220, 2019. 1

Liu, M., Yuan, Z., Ying, Y., and Yang, T. Stochastic AUC maximization with deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=HJepXaVYDr. 1

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018. 1

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of gans. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1, 3.1

Mokhtari, A., Ozdaglar, A. E., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. URL https://api.semanticscholar.org/CorpusID:59222714. 5.3, A

Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/1aa48fc4880bb0c9b8a3bf979d3b917e-Paper.pdf. 3

Park, J. and Ryu, E. K. Exact optimal accelerated complexity for fixed-point iterations. In *International Conference on Machine Learning (ICML)*, pp. 17420–17457. PMLR, 2022. 1

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7.2, G.3

Popov, L. D. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28:845–848, 1980. 1, 5.3, A

Ramirez, J., Sukumaran, R., Bertrand, Q., and Gidel, G. Omega: Optimistic ema gradients. In *International Conference on Machine Learning (ICML)*, 2023. G.2

Ryu, E. K., Taylor, A. B., Bergeling, C., and Giselsson, P. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020. doi: 10.1137/19M1304854. URL https://doi.org/10.1137/19M1304854. 8, H

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-. 1

Taylor, A., Van Scoy, B., and Lessard, L. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning (ICML)*, 2018. 2.3

von Neumann, J. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. doi: 10.1007/BF01448847. URL https://doi.org/10.1007/BF01448847. 1

Ying, Y., Wen, L., and Lyu, S. Stochastic online AUC maximization. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016. 1

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

Yoon, T. and Ryu, E. K. Accelerated minimax algorithms flock together. *arXiv preprint arXiv:2205.11093*, 2022. 1

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep

learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 7.2, G.3

Yu, Y., Lin, T., Mazumdar, E. V., and Jordan, M. Fast distributionally robust learning with variance-reduced min-max optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1219–1250. PMLR, 2022. 1

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *IEEE International Conference on Computer Vision (ICCV(*, pp. 3040–3049, 2021. 1

Zamani, M., Abbaszadehpeivasti, H., and de Klerk, E. Convergence rate analysis of the gradient descent-ascent method for convex-concave saddle-point problems, 2022. 3.1

Zhang, F. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006. E.1

Zhang, G. and Yu, Y. Convergence of gradient methods on bilinear zero-sum games. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlVY04FwH. 6.1, A, G.2, G.2

Zhang, G., Wang, Y., Lessard, L., and Grosse, R. B. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022. 1, 2.1, 3.1, 4.1, 7.1, G.1

Supplementary Material

A. Comparison with OGD

Here we compare **Alex-GDA** to the Optimistic Gradient Descent (**OGD**) method (Popov, 1980), an algorithm based on *simultaneous* updates of the form:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - 2\alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) + \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}), \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + 2\beta\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) - \beta\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}). \end{aligned} \tag{6}$$

We remark that **OGD** takes the same amount of gradient computation as **Sim-GDA**, **Alt-GDA**, and **Alex-GDA**. One may observe that **Alex-GDA** stores the previous *iterates* $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ to compute $\tilde{\boldsymbol{x}}_{k+1}$ and $\tilde{\boldsymbol{y}}_{k+1}$, whereas the implementation of **OGD** requires storing the previous *gradients* $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1})$ and $\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1})$ instead. As a result, while these two algorithms exploit different types of information, the memory consumption of **Alex-GDA** and **OGD** are identical.

As **EG**, it is also known that **OGD** converges with iteration complexity $\tilde{\mathcal{O}}(\kappa)$, where $\kappa = \frac{\max\{L_x, L_y, L_{xy}\}}{\min\{\mu_x, \mu_y\}}$ (Mokhtari et al., 2019). We show that the iteration complexity cannot be strictly better than **Alex-GDA** through the following proposition.

Proposition A.1. *There exists a* 6*-dimensional function* $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ *with* $d_x = d_y = 3$ *such that for any constant step sizes* $\alpha, \beta > 0$*, the convergence of* ***OGD*** *requires an iteration complexity of rate at least*

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log\frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$.

We prove Proposition A.1 in Appendix F.

We also compare **Alex-GDA** with **OGD** in terms of bilinear problem, based upon the analysis of Zhang & Yu (2020). From their results, the iteration complexity of **OGD** is translated to $\tilde{\mathcal{O}}\left(\left(L_{xy}/\mu_{xy}\right)^2\right)$; it matches to the iteration complexity of **Alex-GDA** up to constant factor. In more detail, the authors proved that optimal convergence rate exponent of **OGD** of the form in Equation (6) is approximately $1 - \frac{\mu_{xy}^2}{6L_{xy}^2}$, which is 6 times slower than our optimal rate exponent $\sqrt{\frac{L_{xy}^2 - \mu_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}} \approx 1 - \frac{\mu_{xy}^2}{L_{xy}^2}$ proved in Theorem 6.2. On the other hand, Zhang & Yu (2020) also proved that the alternating variant of **OGD**, *i.e.*, Gauss-Siedel OGD (**GS-OGD**), has an optimal convergence rate exponent $\sqrt{\frac{L_{xy}^2 - \mu_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}}$. It exactly matches our optimal rate of **Alex-GDA** with $\delta = 1$. These facts again buttress our claim that alternating updates are beneficial in minimax optimization.

B. Proofs used in Section 3

Here we prove all theorems related to **Sim-GDA** presented in Section 3.

- In Appendix B.1 we prove Theorem 3.1 which yields a contraction inequality for **Sim-GDA**.
- In Appendix B.2 we prove Corollary 3.2 which derives the corresponding iteration complexity upper bound.
- In Appendix B.3 we prove Theorem 3.3 which yields a matching lower bound for **Sim-GDA**.
- In Appendix B.4 we prove technical propositions and lemmas used throughout the proofs in Appendix B.

B.1. Proof of Theorem 3.1

Here we prove Theorem 3.1 of Section 3, restated below for the sake of readability.

Theorem 3.1. *Suppose that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$. Then, there exists a pair of step sizes α, β with*

$$\alpha\mu_x = \beta\mu_y = \Theta\left(\frac{1}{\kappa_x + \kappa_y + \kappa_{xy}^2}\right),$$

*such that **Sim-GDA** satisfies $\Psi_{k+1}^{\text{Sim}} \leq r\Psi_k^{\text{Sim}}$ with*

$$r = \left(\frac{\left(\kappa_{xy} + \sqrt{\max\{\kappa_x, \kappa_y\} + \kappa_{xy}^2}\right)^2 - 1}{\left(\kappa_{xy} + \sqrt{\max\{\kappa_x, \kappa_y\} + \kappa_{xy}^2}\right)^2 + 1}\right)^2. \tag{4}$$

Proof. Recall that we define the Lyapunov function as

$$\Psi_k^{\text{Sim}} = \frac{1}{\alpha}\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_k - \boldsymbol{y}_\star\|^2.$$

Now we will show that $\Psi_1^{\text{Sim}} \leq \gamma\Psi_0^{\text{Sim}}$ for any choice of initialization points $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ (*i.e.*, set $k = 0$ W.L.O.G.), which directly implies $\Psi_{k+1}^{\text{Sim}} \leq \gamma\Psi_k^{\text{Sim}}$ for all k. Proposition B.1 yields a one-step contraction inequality that applies to **Sim-GDA** with $\alpha < \frac{1}{L_x}$ and $\beta < \frac{1}{L_y}$, *i.e.*, when the step sizes are small enough.

Proposition B.1. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$, **Sim-GDA** with step sizes $\alpha < \frac{1}{L_x}$ and $\beta < \frac{1}{L_y}$ satisfies*

$$\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 \leq r\left(\frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2\right),$$

where the contraction factor is given by

$$r = \max\left\{\left\|\begin{bmatrix} 1 - \alpha L_x & -\sqrt{\alpha\beta}L_{xy} \\ \sqrt{\alpha\beta}L_{xy} & 1 - \beta\mu_y \end{bmatrix}\right\|^2, \left\|\begin{bmatrix} 1 - \alpha\mu_x & -\sqrt{\alpha\beta}L_{xy} \\ \sqrt{\alpha\beta}L_{xy} & 1 - \beta L_y \end{bmatrix}\right\|^2\right\}.$$

We prove Proposition B.1 in Appendix B.4.1.

To find the right step sizes, we search among α, β which satisfies $\alpha/\beta = \mu_y/\mu_x$. This allows us to reduce the problem to optimizing the choice of ζ, which can be defined as

$$\zeta = \alpha\mu_x = \beta\mu_y.$$

Then the contraction factor can be rewritten as

$$r = \max\left\{\left\|\begin{bmatrix} 1 - \zeta\kappa_x & -\zeta\kappa_{xy} \\ \zeta\kappa_{xy} & 1 - \zeta \end{bmatrix}\right\|^2, \left\|\begin{bmatrix} 1 - \zeta & -\zeta\kappa_{xy} \\ \zeta\kappa_{xy} & 1 - \zeta\kappa_y \end{bmatrix}\right\|^2\right\}.$$

For $\kappa \geq 1$, let us define the function $f_\kappa : (0, \infty) \to (0, \infty)$ as:

$$f_\kappa(\zeta) = \left\| \begin{bmatrix} 1 - \zeta\kappa & -\zeta\kappa_{xy} \\ \zeta\kappa_{xy} & 1 - \zeta \end{bmatrix} \right\| = \frac{\kappa - 1}{2} \cdot \zeta + \sqrt{\left(1 - \frac{\kappa + 1}{2} \cdot \zeta\right)^2 + \zeta^2\kappa_{xy}^2}. \tag{7}$$

Then we can simplify as follows:

$$r = \max\left\{ \left(f_{\kappa_x}(\zeta)\right)^2, \left(f_{\kappa_y}(\zeta)\right)^2 \right\}.$$

Proposition B.2 characterizes the optimal choice of ζ and the optimal function value of $f_\kappa(\zeta)$ defined as in (7).

Proposition B.2. *For $f_\kappa : (0, \infty) \to (0, \infty)$ defined as in (7), the minimizer $\zeta^\star$ is equal to*

$$\zeta^\star = \frac{1}{\sqrt{\kappa + \kappa_{xy}^2}} \cdot \frac{2\left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)}{1 + \left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)^2}$$

and the minimum value of f_κ attained at $\zeta^\star$ is equal to

$$f_\kappa(\zeta^\star) = \frac{\left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)^2 - 1}{\left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)^2 + 1}.$$

Moreover, we have $f_{\kappa_x}(\zeta) \geq f_{\kappa_y}(\zeta)$ for all $\zeta \in (0, \infty)$ if and only if $\kappa_x \geq \kappa_y$.

We prove Proposition B.2 in Appendix B.4.2.

If $\kappa_x \geq \kappa_y$, we choose α, β such that

$$\alpha\mu_x = \beta\mu_y = \zeta_x^\star := \frac{1}{\sqrt{\kappa_x + \kappa_{xy}^2}} \cdot \frac{2\left(\kappa_{xy} + \sqrt{\kappa_x + \kappa_{xy}^2}\right)}{1 + \left(\kappa_{xy} + \sqrt{\kappa_x + \kappa_{xy}^2}\right)^2}.$$

Note that $\zeta_x^\star = \Theta\left(\frac{1}{\kappa_x + \kappa_{xy}^2}\right)$. Then, since $f_{\kappa_x}(\zeta) \geq f_{\kappa_y}(\zeta)$, we have

$$r = \max\left\{ \left(f_{\kappa_x}(\zeta_x^\star)\right)^2, \left(f_{\kappa_y}(\zeta_x^\star)\right)^2 \right\} = \left(f_{\kappa_x}(\zeta_x^\star)\right)^2 = \left(\frac{\left(\kappa_{xy} + \sqrt{\kappa_x + \kappa_{xy}^2}\right)^2 - 1}{\left(\kappa_{xy} + \sqrt{\kappa_x + \kappa_{xy}^2}\right)^2 + 1} \right)^2$$

which is identical to (4) when $\kappa_x \geq \kappa_y$.

Similarly, if $\kappa_x < \kappa_y$, we choose α, β such that

$$\alpha\mu_x = \beta\mu_y = \zeta_y^\star := \frac{1}{\sqrt{\kappa_y + \kappa_{xy}^2}} \cdot \frac{2\left(\kappa_{xy} + \sqrt{\kappa_y + \kappa_{xy}^2}\right)}{1 + \left(\kappa_{xy} + \sqrt{\kappa_y + \kappa_{xy}^2}\right)^2}.$$

Note that $\zeta_y^\star = \Theta\left(\frac{1}{\kappa_y + \kappa_{xy}^2}\right)$. Then, since $f_{\kappa_y}(\zeta) \leq f_{\kappa_y}(\zeta)$, we have

$$r = \max\left\{ \left(f_{\kappa_x}(\zeta_y^\star)\right)^2, \left(f_{\kappa_y}(\zeta_y^\star)\right)^2 \right\} = \left(f_{\kappa_y}(\zeta_y^\star)\right)^2 = \left(\frac{\left(\kappa_{xy} + \sqrt{\kappa_y + \kappa_{xy}^2}\right)^2 - 1}{\left(\kappa_{xy} + \sqrt{\kappa_y + \kappa_{xy}^2}\right)^2 + 1} \right)^2$$

which is identical to (4) when $\kappa_x < \kappa_y$. Note that for either case, we have that

$$\alpha\mu_x = \beta\mu_y = \Theta\left(\frac{1}{\max\{\kappa_x, \kappa_y\} + \kappa_{xy}}\right) = \Theta\left(\frac{1}{\kappa_x + \kappa_y + \kappa_{xy}}\right),$$

which concludes the proof.[4] □

B.2. Proof of Corollary 3.2

Here we prove Corollary 3.2 of Section 3, restated below for the sake of readability.

Corollary 3.2. *For the step sizes given as in Theorem 3.1, **Sim-GDA** linearly converges with iteration complexity*

$$\mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right) \cdot \log \frac{\Psi_0^{\text{Sim}}}{A^{\text{Sim}}\epsilon}\right),$$

where $A^{\text{Sim}} = \min\left\{\frac{1}{\alpha}, \frac{1}{\beta}\right\}$.

Proof. Let us define $\xi := \kappa_{xy} + \sqrt{\max\{\kappa_x, \kappa_y\} + \kappa_{xy}^2}$ so that $r = \left(\frac{\xi^2-1}{\xi^2+1}\right)^2$ by Theorem 3.1. By definition we have $\xi^2 = \Theta\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right)$ and $\xi \geq 1$, which gives us

$$\frac{1}{1-r} = \frac{1}{1 - \left(\frac{\xi^2-1}{\xi^2+1}\right)^2} = \frac{\left(\xi^2+1\right)^2}{\left(\xi^2+1\right)^2 - \left(\xi^2-1\right)^2} = \frac{1}{4}\left(\xi + \frac{1}{\xi}\right)^2 = \Theta\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right).$$

Therefore it is sufficient to run

$$K = \mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right) \cdot \log \frac{\Psi_0^{\text{Sim}}}{A^{\text{Sim}}\epsilon}\right)$$

iterations to ensure that $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$, where $A^{\text{Sim}} = \min\left\{\frac{1}{\alpha}, \frac{1}{\beta}\right\}$. □

Remark. Here we present a simpler proof of Corollary 3.2 we discovered afterwards. The proof can achieve the same iteration complexity upper bound with a similar yet slightly different choice of step sizes α, β. Compared to the one using Theorem 3.1, this proof does not require complicated matrix analyses and better extends to algorithms with alternating updates, such as **Alt-GDA** (as in Proposition C.2) or **Alex-GDA** (as in Proposition D.2).

Step 1. Contraction Inequality

We first prove the following proposition.

Proposition B.3. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and **Sim-GDA** with step sizes $\alpha \leq \frac{1}{2L_x}$ and $\beta \leq \frac{1}{2L_y}$, we have*

$$\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 \leq \left(\frac{1}{\alpha} - \mu_x + 2\beta L_{xy}^2\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \left(\frac{1}{\beta} - \mu_y + 2\alpha L_{xy}^2\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 \tag{8}$$

for all $\boldsymbol{x}_0 \in \mathbb{R}^{d_x}, \boldsymbol{y}_0 \in \mathbb{R}^{d_y}$.

Proof. Recall that **Sim-GDA** takes updates of the form:

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0),$$
$$\boldsymbol{y}_1 = \boldsymbol{y}_0 + \beta\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0, \boldsymbol{y}_0).$$

[4] Note that for $a, b \geq 0$, we have $\max\{a, b\} = \Theta(a + b)$ since $\frac{a+b}{2} \leq \max\{a, b\} \leq a + b$.

From this, we can deduce that

$$\begin{aligned}\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 &= \frac{1}{\alpha}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+\frac{2}{\alpha}\left\langle \boldsymbol{x}_1-\boldsymbol{x}_0,\boldsymbol{x}_0-\boldsymbol{x}_\star\right\rangle+\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2\\ &= \frac{1}{\alpha}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2-2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0),\boldsymbol{x}_0-\boldsymbol{x}_\star\right\rangle+\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2,\\ \frac{1}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2 &= \frac{1}{\beta}\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+\frac{2}{\beta}\left\langle \boldsymbol{y}_1-\boldsymbol{y}_0,\boldsymbol{y}_0-\boldsymbol{y}_\star\right\rangle+\frac{1}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_0\|^2\\ &= \frac{1}{\beta}\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+2\left\langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0,\boldsymbol{y}_0),\boldsymbol{y}_0-\boldsymbol{y}_\star\right\rangle+\beta\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2.\end{aligned}$$

Note that μ_x-strong convexity of $f(\cdot,\boldsymbol{y}_0)$ yields

$$-2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0),\boldsymbol{x}_0-\boldsymbol{x}_\star\right\rangle\le-\mu_x\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2-2(f(\boldsymbol{x}_0,\boldsymbol{y}_0)-f(\boldsymbol{x}_\star,\boldsymbol{y}_0)),\tag{9}$$

and μ_y-strong concavity of $f(\boldsymbol{x}_0,\cdot)$ yields

$$2\left\langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0,\boldsymbol{y}_0),\boldsymbol{y}_0-\boldsymbol{y}_\star\right\rangle\le-\mu_y\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2-2(f(\boldsymbol{x}_0,\boldsymbol{y}_\star)-f(\boldsymbol{x}_0,\boldsymbol{y}_0)).\tag{10}$$

Moreover, since f is convex-concave and has Lipschitz gradients[5], we have

$$-2(f(\boldsymbol{x}_0,\boldsymbol{y}_\star)-f(\boldsymbol{x}_\star,\boldsymbol{y}_\star))\le-\frac{1}{L_x}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_\star)\|^2,\tag{11}$$

$$-2(f(\boldsymbol{x}_\star,\boldsymbol{y}_\star)-f(\boldsymbol{x}_\star,\boldsymbol{y}_0))\le-\frac{1}{L_y}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_0)\|^2.\tag{12}$$

Applying (9)–(12), we have

$$\begin{aligned}&\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+\frac{1}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2\\ &\le\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2\\ &+\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2+\beta\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2-\frac{1}{L_x}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_\star)\|^2-\frac{1}{L_y}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_0)\|^2.\end{aligned}$$

If $\alpha\le\frac{1}{2L_x}$ and $\beta\le\frac{1}{2L_y}$, we can use the triangle inequality and the Lipschitz gradient condition for L_{xy} to obtain

$$\begin{aligned}\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2-\frac{1}{L_x}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_\star)\|^2 &\le\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2-2\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_\star)\|^2\\ &\le 2\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)-\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_\star)\|^2\\ &\le 2\alpha L_{xy}^2\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2,\\ \beta\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2-\frac{1}{L_y}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_0)\|^2 &\le\beta\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2-2\beta\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_0)\|^2\\ &\le 2\beta\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)-\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_0)\|^2\\ &\le 2\beta L_{xy}^2\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2,\end{aligned}$$

which boils down to (8). □

STEP 2. ITERATION COMPLEXITY

Now let us show that Proposition B.3 can guarantee the same iteration complexity as in Corollary 3.2 when

$$\alpha=\frac{1}{2}\cdot\min\left\{\frac{1}{L_x},\frac{\mu_y}{2L_{xy}^2}\right\},\quad\beta=\frac{1}{2}\cdot\min\left\{\frac{1}{L_y},\frac{\mu_x}{2L_{xy}^2}\right\}.$$

[5]Note that the Lipschitz gradient conditions for L_x and L_y are equivalent to the widely used notion of *smoothness* in convex optimization literature.

Proof. If $\alpha \leq \frac{\mu_y}{4L_{xy}^2}$ and $\beta \leq \frac{\mu_x}{4L_{xy}^2}$, Proposition B.3 implies

$$\begin{aligned}\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 &\leq \left(\frac{1}{\alpha} - \mu_x + 2\beta L_{xy}^2\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \left(\frac{1}{\beta} - \mu_y + 2\alpha L_{xy}^2\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 \\ &\leq \left(\frac{1}{\alpha} - \frac{\mu_x}{2}\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \left(\frac{1}{\beta} - \frac{\mu_y}{2}\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2.\end{aligned}$$

Hence we have $\Psi_1^{\text{Sim}} \leq r\Psi_0^{\text{Sim}}$ for $r = \max\{1 - \alpha\mu_x/2, 1 - \beta\mu_y/2\}$, and

$$\begin{aligned}\frac{1}{1-r} \leq \max\left\{\frac{1}{\alpha\mu_x}, \frac{1}{\beta\mu_y}\right\} &= \max\left\{\Theta\left(\kappa_x + \kappa_{xy}^2\right), \Theta\left(\kappa_y + \kappa_{xy}^2\right)\right\} \\ &= \Theta\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right).\end{aligned}$$

Therefore it is sufficient to take

$$K = \mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right) \cdot \log \frac{\Psi_0^{\text{Sim}}}{A^{\text{Sim}}\epsilon}\right)$$

iterations to ensure that $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$, where $A^{\text{Sim}} = \min\left\{\frac{1}{\alpha}, \frac{1}{\beta}\right\}$. □

B.3. Proof of Theorem 3.3

Here we prove Theorem 3.3 of Section 3, restated below for the sake of readability.

Theorem 3.3. *There exists a 6-dimensional function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ with $d_x = d_y = 3$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of **Sim-GDA** requires an iteration complexity of rate at least*

$$\Omega\left(\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$.

Proof. We construct the worst-case function as follows:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\begin{bmatrix} x \\ s \\ t \\ y \\ u \\ v \end{bmatrix}^\top \begin{bmatrix} \mu_x & 0 & 0 & L_{xy} & 0 & 0 \\ 0 & \mu_x & 0 & 0 & 0 & 0 \\ 0 & 0 & L_x & 0 & 0 & 0 \\ L_{xy} & 0 & 0 & -\mu_y & 0 & 0 \\ 0 & 0 & 0 & 0 & -\mu_y & 0 \\ 0 & 0 & 0 & 0 & 0 & -L_y \end{bmatrix} \begin{bmatrix} x \\ s \\ t \\ y \\ u \\ v \end{bmatrix},$$

where $\boldsymbol{x} = (x, s, t)$ and $\boldsymbol{y} = (y, u, v)$. We can easily check that f is a quadratic function (*i.e.,* the Hessian is constant) such that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and $\boldsymbol{x}_\star = \boldsymbol{y}_\star = \boldsymbol{0} \in \mathbb{R}^3$.

As a first step, we will find a set of necessary conditions on step sizes for convergence, and then compute (at least) how large the number of iterations K of **Sim-GDA** we need to accomplish $\|\boldsymbol{x}_K\|^2 + \|\boldsymbol{y}_K\|^2 < \epsilon$. To this end, we first observe that the k-th step of **Sim-GDA** satisfies

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 - \alpha\mu_x & -\alpha L_{xy} \\ \beta L_{xy} & 1 - \beta\mu_y \end{bmatrix}}_{\triangleq \boldsymbol{P}} \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \tag{13}$$

$$s_{k+1} = (1 - \alpha\mu_x)s_k, \tag{14}$$

$$t_{k+1} = (1 - \alpha L_x)t_k, \tag{15}$$

$$u_{k+1} = (1 - \beta\mu_y)u_k, \tag{16}$$

$$v_{k+1} = (1 - \beta L_y)v_k. \tag{17}$$

To assure the convergence of iterations (15) and (17), the step sizes α and β are required to be

$$\alpha < \frac{2}{L_x} \quad \text{and} \quad \beta < \frac{2}{L_y}. \tag{18}$$

Also, to guarantee $\|\boldsymbol{x}_K\|^2 + \|\boldsymbol{y}_K\|^2 < \epsilon$, we need from (14) and (16) that $s_K^2 < \mathcal{O}(\epsilon)$ and $u_K^2 < \mathcal{O}(\epsilon)$, respectively. These two necessary conditions require an iteration number of at least:

$$K = \Omega\left(\left(\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y}\right) \cdot \log\frac{1}{\epsilon}\right). \tag{19}$$

Note that (18) automatically yields

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y). \tag{20}$$

From now on, we deal with the remaining proof case by case with respect to the step sizes α and β.

Case 1. Suppose that α and β satisfies $\left(\frac{\alpha\mu_x - \beta\mu_y}{2}\right)^2 \leq \alpha\beta L_{xy}^2$, which is equivalent to the eigenvalues of the matrix $\boldsymbol{P}$ defined in Equation (13) being complex. We can check that, for $i = \sqrt{-1}$, the eigenvalues of $\boldsymbol{P}$ can be expressed as

$$\lambda = 1 - \frac{\alpha\mu_x + \beta\mu_y}{2} \pm i\sqrt{\alpha\beta L_{xy}^2 - \left(\frac{\alpha\mu_x - \beta\mu_y}{2}\right)^2}.$$

We recall a well-known convergence theory of matrix iteration in Proposition B.4.

Proposition B.4 (Horn & Johnson (2012), Theorem 5.6.12, Corollary 5.6.13)**.** *For a square matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ and a sequence of m-dimensional vectors $(\boldsymbol{v}_k)$, the matrix iteration $\boldsymbol{v}_{k+1} = \boldsymbol{A}\boldsymbol{v}_k$ converges as $\boldsymbol{v}_k \to \boldsymbol{0}$ with arbitrarily chosen initialization $\boldsymbol{v}_0$ if and only if the spectral radius $\rho(\boldsymbol{A})$ of $\boldsymbol{A}$ is less than 1. In this case, the convergence rate is written as $O((\rho(\boldsymbol{A}) + \epsilon)^k)$, where ϵ is an any given positive number.*

Noting that

$$\rho(\boldsymbol{P})^2 = \left(1 - \frac{\alpha\mu_x + \beta\mu_y}{2}\right)^2 + \alpha\beta L_{xy}^2 - \left(\frac{\alpha\mu_x - \beta\mu_y}{2}\right)^2 = 1 - (\alpha\mu_x + \beta\mu_y) + \alpha\beta\left(\mu_x\mu_y + L_{xy}^2\right),$$

in order to assure convergence of iteration (13), we need

$$\rho(\boldsymbol{P})^2 < 1 \iff \beta < \frac{\mu_x + r\mu_y}{\mu_x\mu_y + L_{xy}^2} \iff \alpha < \frac{\frac{1}{r}\mu_x + \mu_y}{\mu_x\mu_y + L_{xy}^2},$$

where $r = \frac{\beta}{\alpha}$ is the ratio of step sizes. Combined with (18), we have

$$\frac{1}{\alpha\mu_x} > \max\left\{\frac{L_x}{2\mu_x}, \frac{rL_y}{2\mu_x}, \frac{\mu_x\mu_y + L_{xy}^2}{\frac{1}{r}\mu_x^2 + \mu_x\mu_y}\right\}, \tag{21}$$

$$\frac{1}{\beta\mu_y} > \max\left\{\frac{L_x}{2r\mu_y}, \frac{L_y}{2\mu_y}, \frac{\mu_x\mu_y + L_{xy}^2}{\mu_x\mu_y + r\mu_y^2}\right\}. \tag{22}$$

If $r \geq \frac{\mu_x}{\mu_y}$, then by (21), $\frac{1}{\alpha\mu_x} = \Omega\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right)$. On the other hand, if $r < \frac{\mu_x}{\mu_y}$, then by (22), $\frac{1}{\beta\mu_y} = \Omega\left(\kappa_x + \kappa_y + \kappa_{xy}^2\right)$. Therefore, we have a desired lower bound of iteration complexity for the first case, deduced from (19).

Case 2. Suppose that α and β satisfies $\left(\frac{\alpha\mu_x-\beta\mu_y}{2}\right)^2 > \alpha\beta L_{xy}^2$. Note that this is equivalent to

$$\left|\sqrt{\frac{\alpha\mu_x}{\beta\mu_y}}-\sqrt{\frac{\beta\mu_y}{\alpha\mu_x}}\right| > 2\kappa_{xy}. \tag{23}$$

If $r \geq \frac{\mu_x}{\mu_y}$, *i.e.*, $\frac{\alpha\mu_x}{\beta\mu_y} \leq \frac{\beta\mu_y}{\alpha\mu_x}$, then it implies $\frac{\beta\mu_y}{\alpha\mu_x} > 4\kappa_{xy}^2$. Thus, combined with (20), we have

$$\frac{1}{\alpha\mu_x}+\frac{1}{\beta\mu_y} = \frac{1}{2}\cdot\frac{1}{\alpha\mu_x}+\frac{1}{\beta\mu_y}\left(\frac{1}{2}\cdot\frac{\beta\mu_y}{\alpha\mu_x}+1\right) = \Omega\left(\kappa_x+\kappa_y(\kappa_{xy}^2+1)\right) = \Omega\left(\kappa_x+\kappa_y+\kappa_{xy}^2\right).$$

On the other hand, if $r < \frac{\mu_x}{\mu_y}$, *i.e.*, $\frac{\alpha\mu_x}{\beta\mu_y} > \frac{\beta\mu_y}{\alpha\mu_x}$, then it implies $\frac{\alpha\mu_x}{\beta\mu_y} > 4\kappa_{xy}^2$. Thus, combined with (20), we have

$$\frac{1}{\alpha\mu_x}+\frac{1}{\beta\mu_y} = \frac{1}{\alpha\mu_x}\left(1+\frac{1}{2}\cdot\frac{\alpha\mu_x}{\beta\mu_y}\right)+\frac{1}{2}\cdot\frac{1}{\beta\mu_y} = \Omega\left(\kappa_x(1+\kappa_{xy}^2)+\kappa_y\right) = \Omega\left(\kappa_x+\kappa_y+\kappa_{xy}^2\right).$$

Therefore, from (19) we can obtain the desired lower bound for the second case as well, which concludes the proof. □

B.4. Proofs used in Appendix B

Here we prove some technical propositions and lemmas used throughout Appendix B.

B.4.1. Proof of Proposition B.1

Here we prove Proposition B.1, restated below for the sake of readability.

Proposition B.1. *For $f \in \mathcal{F}(\mu_x,\mu_y,L_x,L_y,L_{xy})$, **Sim-GDA** with step sizes $\alpha < \frac{1}{L_x}$ and $\beta < \frac{1}{L_y}$ satisfies*

$$\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+\frac{1}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2 \leq r\left(\frac{1}{\alpha}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+\frac{1}{\beta}\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2\right),$$

where the contraction factor is given by

$$r = \max\left\{\left\|\begin{bmatrix}1-\alpha L_x & -\sqrt{\alpha\beta}L_{xy}\\ \sqrt{\alpha\beta}L_{xy} & 1-\beta\mu_y\end{bmatrix}\right\|^2, \left\|\begin{bmatrix}1-\alpha\mu_x & -\sqrt{\alpha\beta}L_{xy}\\ \sqrt{\alpha\beta}L_{xy} & 1-\beta L_y\end{bmatrix}\right\|^2\right\}.$$

Proof. Recall that **Sim-GDA** takes updates of the form:

$$\begin{aligned}\boldsymbol{x}_1 &= \boldsymbol{x}_0-\alpha\nabla_{\boldsymbol{x}}f(\boldsymbol{x}_0,\boldsymbol{y}_0),\\ \boldsymbol{y}_1 &= \boldsymbol{y}_0+\beta\nabla_{\boldsymbol{y}}f(\boldsymbol{x}_0,\boldsymbol{y}_0).\end{aligned} \tag{24}$$

For simplicity, let us denote $\boldsymbol{z} = \begin{bmatrix}\boldsymbol{x}^\top & \boldsymbol{y}^\top\end{bmatrix}^\top \in \mathbb{R}^{d_x+d_y}$, and define

$$\nu(\boldsymbol{z}) := \begin{bmatrix}\nabla_{\boldsymbol{x}}f(\boldsymbol{z})\\ -\nabla_{\boldsymbol{y}}f(\boldsymbol{z})\end{bmatrix}.$$

For instance, $\boldsymbol{z}_0 = \begin{bmatrix}\boldsymbol{x}_0^\top & \boldsymbol{y}_0^\top\end{bmatrix}^\top$ and $\boldsymbol{z}_\star = \begin{bmatrix}\boldsymbol{x}^{\star\top} & \boldsymbol{y}^{\star\top}\end{bmatrix}^\top$.

Let us define matrices $\boldsymbol{A} \in \mathbb{R}^{d_x\times d_x}$, $\boldsymbol{B} \in \mathbb{R}^{d_x\times d_y}$, and $\boldsymbol{C} \in \mathbb{R}^{d_y\times d_y}$ as

$$\boldsymbol{A} := \int_0^1 \nabla_{\boldsymbol{xx}}^2 f(t\boldsymbol{z}_0+(1-t)\boldsymbol{z}_\star)dt,\ \ \boldsymbol{B} := \int_0^1 \nabla_{\boldsymbol{xy}}^2 f(t\boldsymbol{z}_0+(1-t)\boldsymbol{z}_\star)dt,\ \ \boldsymbol{C} := -\int_0^1 \nabla_{\boldsymbol{yy}}^2 f(t\boldsymbol{z}_0+(1-t)\boldsymbol{z}_\star)dt.$$

Since $f \in \mathcal{F}(\mu_x,\mu_y,L_x,L_y,L_{xy})$, we have $\mu_x\boldsymbol{I} \preceq \boldsymbol{A} \preceq L_x\boldsymbol{I}$, $\mu_y\boldsymbol{I} \preceq \boldsymbol{C} \preceq L_y\boldsymbol{I}$, and $\|\boldsymbol{B}\| \leq L_{xy}$.

Also, by chain rule, we have the following identities:

$$\begin{aligned}\nabla_{\boldsymbol{x}}f(\boldsymbol{x}_0,\boldsymbol{y}_0) &= \boldsymbol{A}(\boldsymbol{x}_0-\boldsymbol{x}_\star)+\boldsymbol{B}(\boldsymbol{y}_0-\boldsymbol{y}_\star),\\ \nabla_{\boldsymbol{y}}f(\boldsymbol{x}_0,\boldsymbol{y}_0) &= \boldsymbol{B}^\top(\boldsymbol{x}_0-\boldsymbol{x}_\star)-\boldsymbol{C}(\boldsymbol{y}_0-\boldsymbol{y}_\star).\end{aligned}$$

For simplicity, we assume W.L.O.G. $\boldsymbol{x}_\star = \mathbf{0}$ $(\in \mathbb{R}^{d_x})$ and $\boldsymbol{y}_\star = \mathbf{0}$ $(\in \mathbb{R}^{d_y})$. Then we have

$$\begin{aligned}\begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_1\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_1\end{bmatrix} &= \begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_0\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_0\end{bmatrix} - \begin{bmatrix}\sqrt{\alpha}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\\ -\sqrt{\beta}\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_0,\boldsymbol{y}_0)\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_0\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_0\end{bmatrix} - \begin{bmatrix}\sqrt{\alpha}(\boldsymbol{A}\boldsymbol{x}_0+\boldsymbol{B}\boldsymbol{y}_0)\\ -\sqrt{\beta}(\boldsymbol{B}^\top\boldsymbol{x}_0-\boldsymbol{C}\boldsymbol{y}_0)\end{bmatrix}\\ &= \begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_0\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_0\end{bmatrix} - \begin{bmatrix}\alpha\boldsymbol{A} & \sqrt{\alpha\beta}\boldsymbol{B}\\ -\sqrt{\alpha\beta}\boldsymbol{B}^\top & \beta\boldsymbol{C}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_0\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{I}-\alpha\boldsymbol{A} & -\sqrt{\alpha\beta}\boldsymbol{B}\\ \sqrt{\alpha\beta}\boldsymbol{B}^\top & \boldsymbol{I}-\beta\boldsymbol{C}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_0\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_0\end{bmatrix}.\end{aligned}$$

This means that it is enough to show that

$$\left\|\begin{bmatrix}\boldsymbol{I}-\alpha\boldsymbol{A} & -\sqrt{\alpha\beta}\boldsymbol{B}\\ \sqrt{\alpha\beta}\boldsymbol{B}^\top & \boldsymbol{I}-\beta\boldsymbol{C}\end{bmatrix}\right\|^2 \le r = \max\left\{\left\|\begin{bmatrix}1-\alpha L_x & -\sqrt{\alpha\beta}L_{xy}\\ \sqrt{\alpha\beta}L_{xy} & 1-\beta\mu_y\end{bmatrix}\right\|^2, \left\|\begin{bmatrix}1-\alpha\mu_x & -\sqrt{\alpha\beta}L_{xy}\\ \sqrt{\alpha\beta}L_{xy} & 1-\beta L_y\end{bmatrix}\right\|^2\right\}, \tag{25}$$

since if this is true, then we automatically have

$$\frac{1}{\alpha}\|\boldsymbol{x}_1\|^2+\frac{1}{\beta}\|\boldsymbol{y}_1\|^2 = \left\|\begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_1\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_1\end{bmatrix}\right\|^2 \le r\left\|\begin{bmatrix}\frac{1}{\sqrt{\alpha}}\boldsymbol{x}_0\\ \frac{1}{\sqrt{\beta}}\boldsymbol{y}_0\end{bmatrix}\right\|^2 = r\left(\frac{1}{\alpha}\|\boldsymbol{x}_0\|^2+\frac{1}{\beta}\|\boldsymbol{y}_0\|^2\right).$$

To prove Equation (25), the matrix norm can be bounded via Lemma B.5.

Lemma B.5. *Suppose that $\boldsymbol{X}\in\mathbb{R}^{d_x\times d_x}$, $\boldsymbol{Y}\in\mathbb{R}^{d_y\times d_y}$, $\boldsymbol{W}\in\mathbb{R}^{d_x\times d_y}$ satisfy*

$$t_x\boldsymbol{I}\preceq\boldsymbol{X}\preceq s_x\boldsymbol{I},\ \ t_y\boldsymbol{I}\preceq\boldsymbol{Y}\preceq s_y\boldsymbol{I},\ \ \|\boldsymbol{W}\|\le\ell$$

for some constants $t_x,t_y,s_x,s_y>0$ and $\ell\ge 0$. Then the block matrix $\boldsymbol{M}\in\mathbb{R}^{(d_x+d_y)\times(d_x+d_y)}$ of the form

$$\boldsymbol{M}=\begin{bmatrix}\boldsymbol{X} & -\boldsymbol{W}\\ \boldsymbol{W}^\top & \boldsymbol{Y}\end{bmatrix}$$

satisfies the matrix norm inequality

$$\|\boldsymbol{M}\|\le\max\left\{\left\|\begin{bmatrix}s_x & -\ell\\ \ell & t_y\end{bmatrix}\right\|, \left\|\begin{bmatrix}t_x & -\ell\\ \ell & s_y\end{bmatrix}\right\|\right\}.$$

We prove Lemma B.5 in Appendix B.4.3.

By observing that $1-\alpha L_x>0$, $1-\beta L_y>0$ and

$$(1-\alpha L_x)\boldsymbol{I}\preceq\boldsymbol{I}-\alpha\boldsymbol{A}\preceq(1-\alpha\mu_x)\boldsymbol{I},\ \ (1-\beta L_y)\boldsymbol{I}\preceq\boldsymbol{I}-\beta\boldsymbol{C}\preceq(1-\beta\mu_y)\boldsymbol{I},\ \ \|\sqrt{\alpha\beta}\boldsymbol{B}\|\le\sqrt{\alpha\beta}L_{xy},$$

we can use Lemma B.5 with $\boldsymbol{X}=\boldsymbol{I}-\alpha\boldsymbol{A}$, $\boldsymbol{Y}=\boldsymbol{I}-\beta\boldsymbol{C}$, $\boldsymbol{W}=\sqrt{\alpha\beta}\boldsymbol{B}$, and

$$t_x=1-\alpha L_x,\ \ t_y=1-\beta L_y,\ \ s_x=1-\alpha\mu_x,\ \ s_y=1-\beta\mu_y,\ \ \ell=\sqrt{\alpha\beta}L_{xy}$$

which immediately proves Equation (25), and therefore Proposition B.1. □

B.4.2. Proof of Proposition B.2

Here we prove Proposition B.2, restated below for the sake of readability.

Proposition B.2. *For $f_\kappa:(0,\infty)\to(0,\infty)$ defined as in (7), the minimizer $\zeta^\star$ is equal to*

$$\zeta^\star=\frac{1}{\sqrt{\kappa+\kappa_{xy}^2}}\cdot\frac{2\left(\kappa_{xy}+\sqrt{\kappa+\kappa_{xy}^2}\right)}{1+\left(\kappa_{xy}+\sqrt{\kappa+\kappa_{xy}^2}\right)^2}$$

and the minimum value of f_κ attained at $\zeta^\star$ is equal to

$$f_\kappa(\zeta^\star)=\frac{\left(\kappa_{xy}+\sqrt{\kappa+\kappa_{xy}^2}\right)^2-1}{\left(\kappa_{xy}+\sqrt{\kappa+\kappa_{xy}^2}\right)^2+1}.$$

Moreover, we have $f_{\kappa_x}(\zeta)\ge f_{\kappa_y}(\zeta)$ for all $\zeta\in(0,\infty)$ if and only if $\kappa_x\ge\kappa_y$.

Proof. Recall that we define

$$f_\kappa(\zeta) = \frac{\kappa - 1}{2} \cdot \zeta + \sqrt{\left(1 - \frac{\kappa + 1}{2} \cdot \zeta\right)^2 + \zeta^2 \kappa_{xy}^2}.$$

Then the first two results of the proposition are direct consequences of Lemma B.6.

Lemma B.6. *Suppose that $A, B, C \geq 0$ and $A < B$. Then for the function $f : (0, \infty) \to (0, \infty)$ of the following form:*

$$f(x) = Ax + \sqrt{(1 - Bx)^2 + C^2 x^2},$$

the minimizer is equal to

$$x_\star = \frac{1}{D} \cdot \frac{2(C + D)(B - A)}{(C + D)^2 + (B - A)^2},$$

and the minimum value attained at $x_\star$ is equal to

$$f(x_\star) = \frac{(C + D)^2 - (B - A)^2}{(C + D)^2 + (B - A)^2},$$

where $D = \sqrt{B^2 + C^2 - A^2}$.

We prove Lemma B.6 in Appendix B.4.4.

We can use ζ as x and plug in the following values into Lemma B.6:

$$A = \frac{\kappa - 1}{2}, \;\; B = \frac{\kappa + 1}{2}, \;\; C = \kappa_{xy}, \;\; D = \sqrt{B^2 + C^2 - A^2} = \sqrt{\kappa + \kappa_{xy}^2},$$

which yields

$$B - A = 1, \;\; C + D = \kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}.$$

Then, for the choice

$$\zeta^\star = \frac{1}{D} \cdot \frac{2(C + D)(B - A)}{(C + D)^2 + (B - A)^2} = \frac{1}{\sqrt{\kappa + \kappa_{xy}^2}} \cdot \frac{2\left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)}{1 + \left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)^2},$$

we can obtain the optimal value

$$f_\kappa(\zeta^\star) = \frac{1}{D} \cdot \frac{(C + D)^2 - (B - A)^2}{(C + D)^2 + (B - A)^2} = \frac{\left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)^2 - 1}{\left(\kappa_{xy} + \sqrt{\kappa + \kappa_{xy}^2}\right)^2 + 1}.$$

The last result of the proposition is a direct consequence of the following lemma.

Lemma B.7. *Suppose that $A_1, A_2, B_1, B_2, C \geq 0$ satisfies $A_1 \leq B_1$, $A_2 \leq B_2$, and $A_2 - A_1 = B_2 - B_1 \geq 0$. Then for the functions $f_1, f_2 : (0, \infty) \to (0, \infty)$ of the following form:*

$$f_1(x) = A_1 x + \sqrt{(1 - B_1 x)^2 + C^2 x^2}, \;\; f_2(x) = A_2 x + \sqrt{(1 - B_2 x)^2 + C^2 x^2},$$

we have $f_1(x) \leq f_2(x)$ for all $x > 0$.

We prove Lemma B.7 in Appendix B.4.5.

If $\kappa_x \geq \kappa_y$, we can plug in the following values:

$$A_1 = \frac{\kappa_y - 1}{2}, \;\; A_2 = \frac{\kappa_x - 1}{2}, \;\; B_1 = \frac{\kappa_y + 1}{2}, \;\; B_2 = \frac{\kappa_x + 1}{2}, \;\; C = \kappa_{xy},$$

so that $A_2 - A_1 = B_2 - B_1 = \kappa_x - \kappa_y \geq 0$ and Lemma B.7 implies $f_{\kappa_x}(\zeta) \geq f_{\kappa_y}(\zeta)$.

If $\kappa_x \leq \kappa_y$, we can change orders as:

$$A_1 = \frac{\kappa_x - 1}{2}, \;\; A_2 = \frac{\kappa_y - 1}{2}, \;\; B_1 = \frac{\kappa_x + 1}{2}, \;\; B_2 = \frac{\kappa_y + 1}{2}, \;\; C = \kappa_{xy},$$

so that $A_2 - A_1 = B_2 - B_1 = \kappa_y - \kappa_x \geq 0$ and Lemma B.7 implies $f_{\kappa_x}(\zeta) \leq f_{\kappa_y}(\zeta)$.

Therefore we can conclude that $f_{\kappa_x}(\zeta) \geq f_{\kappa_y}(\zeta)$ for all $\zeta \in (0, \infty)$ if and only if $\kappa_x \geq \kappa_y$. □

B.4.3. Proof of Lemma B.5

Here we prove Lemma B.5, restated below for the sake of readability.

Lemma B.5. *Suppose that $\boldsymbol{X} \in \mathbb{R}^{d_x \times d_x}$, $\boldsymbol{Y} \in \mathbb{R}^{d_y \times d_y}$, $\boldsymbol{W} \in \mathbb{R}^{d_x \times d_y}$ satisfy*

$$t_x \boldsymbol{I} \preceq \boldsymbol{X} \preceq s_x \boldsymbol{I}, \;\; t_y \boldsymbol{I} \preceq \boldsymbol{Y} \preceq s_y \boldsymbol{I}, \;\; \|\boldsymbol{W}\| \leq \ell$$

for some constants $t_x, t_y, s_x, s_y > 0$ and $\ell \geq 0$. Then the block matrix $\boldsymbol{M} \in \mathbb{R}^{(d_x+d_y)\times(d_x+d_y)}$ of the form

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{X} & -\boldsymbol{W} \\ \boldsymbol{W}^\top & \boldsymbol{Y} \end{bmatrix}$$

satisfies the matrix norm inequality

$$\|\boldsymbol{M}\| \leq \max\left\{ \left\| \begin{bmatrix} s_x & -\ell \\ \ell & t_y \end{bmatrix} \right\|, \left\| \begin{bmatrix} t_x & -\ell \\ \ell & s_y \end{bmatrix} \right\| \right\}.$$

Proof. We first observe that the following matrix norms are equal:

$$\left\| \begin{bmatrix} \boldsymbol{X} & -\boldsymbol{W} \\ \boldsymbol{W}^\top & \boldsymbol{Y} \end{bmatrix} \right\| = \Bigg\| \underbrace{\begin{bmatrix} \boldsymbol{X} & \boldsymbol{W} \\ \boldsymbol{W}^\top & -\boldsymbol{Y} \end{bmatrix}}_{\triangleq \boldsymbol{M}'} \Bigg\|.$$

Let $\lambda_{\max}^{\boldsymbol{M}'}$ and $\lambda_{\min}^{\boldsymbol{M}'}$ be the maximum and minimum eigenvalues of $\boldsymbol{M}'$, respectively. Since $\boldsymbol{M}'$ is a symmetric matrix, the matrix norm of $\boldsymbol{M}'$ is equal to

$$\|\boldsymbol{M}'\| = \max\left\{ |\lambda_{\max}^{\boldsymbol{M}'}|, \; |\lambda_{\min}^{\boldsymbol{M}'}| \right\}. \tag{26}$$

Since $\boldsymbol{X} \succ 0$ and $-\boldsymbol{Y} \prec 0$, we can observe that $\boldsymbol{M}'$ is neither positive definite nor negative definite[6], i.e., $\lambda_{\max}^{\boldsymbol{M}'} \geq 0 \geq \lambda_{\min}^{\boldsymbol{M}'}$. Hence we can rewrite:

$$\|\boldsymbol{M}'\| = \max\left\{ \lambda_{\max}^{\boldsymbol{M}'}, \; -\lambda_{\min}^{\boldsymbol{M}'} \right\}. \tag{27}$$

Given a symmetric matrix $\boldsymbol{S} \in \mathbb{S}^d$, we have the following identities:

$$\lambda_{\max}^{\boldsymbol{S}} = \sup_{\boldsymbol{z} \in \mathbb{R}^d, \|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \boldsymbol{S} \boldsymbol{z}, \quad \lambda_{\min}^{\boldsymbol{S}} = \inf_{\boldsymbol{z} \in \mathbb{R}^d, \|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \boldsymbol{S} \boldsymbol{z}, \tag{28}$$

where $\lambda_{\max}^{\boldsymbol{S}}$ and $\lambda_{\min}^{\boldsymbol{S}}$ are the maximum and minimum eigenvalues of $\boldsymbol{S}$, respectively. Moreover, the $\sup$ for $\lambda_{\max}^{\boldsymbol{S}}$ and $\inf$ for $\lambda_{\min}^{\boldsymbol{S}}$ is attained when the unit vector $\boldsymbol{z}$ is aligned with the eigenvectors corresponding to $\lambda_{\max}^{\boldsymbol{S}}$ and $\lambda_{\min}^{\boldsymbol{S}}$.

Now we will show that

$$\lambda_{\max}^{\boldsymbol{M}'} \leq \left\| \begin{bmatrix} s_x & -\ell \\ \ell & t_y \end{bmatrix} \right\| \quad \text{and} \quad -\lambda_{\min}^{\boldsymbol{M}'} \leq \left\| \begin{bmatrix} t_x & -\ell \\ \ell & s_y \end{bmatrix} \right\|. \tag{29}$$

[6] It is easy if we think of the contrapositive— any block partition of a PD matrix must have PD block diagonals.

Maximum Eigenvalue. The maximum eigenvalue of $\boldsymbol{M}'$ is equal to

$$\begin{aligned}\lambda_{\max}^{\boldsymbol{M}'} &= \sup_{\boldsymbol{z}\in\mathbb{R}^{d_x+d_y},\|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \boldsymbol{M}'\boldsymbol{z} \\ &= \sup_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^{d_x},\|\boldsymbol{x}\|=1\\ \boldsymbol{y}\in\mathbb{R}^{d_y},\|\boldsymbol{y}\|=1}} \begin{bmatrix} p\boldsymbol{x}\\ q\boldsymbol{y}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{X} & \boldsymbol{W}\\ \boldsymbol{W}^\top & -\boldsymbol{Y}\end{bmatrix}\begin{bmatrix} p\boldsymbol{x}\\ q\boldsymbol{y}\end{bmatrix} \\ &= \sup_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^{d_x},\|\boldsymbol{x}\|=1\\ \boldsymbol{y}\in\mathbb{R}^{d_y},\|\boldsymbol{y}\|=1}} \left(p^2\boldsymbol{x}^\top\boldsymbol{X}\boldsymbol{x} + 2pq\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{y} - q^2\boldsymbol{y}^\top\boldsymbol{Y}\boldsymbol{y}\right),\end{aligned}$$

where we reparameterize $\boldsymbol{z} = \begin{bmatrix} p\boldsymbol{x}^\top & q\boldsymbol{y}^\top\end{bmatrix}^\top$ such that $\boldsymbol{x}\in\mathbb{R}^{d_x}$, $\boldsymbol{y}\in\mathbb{R}^{d_y}$ satisfies $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$, and $p^2+q^2 = 1$.

First, suppose that $\ell > 0$, *i.e.*, $\boldsymbol{W}\neq 0$. Let $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{W}$, where $\boldsymbol{U} = [\boldsymbol{u}_1,\dots,\boldsymbol{u}_r]\in\mathbb{R}^{d_x\times r}$ and $\boldsymbol{V} = [\boldsymbol{v}_1,\dots,\boldsymbol{v}_r]\in\mathbb{R}^{d_y\times r}$ are matrices with orthonormal columns and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1,\dots,\sigma_r)\in\mathbb{R}^{r\times r}$ is a diagonal matrix with (strictly) positive entries. (Note that $1\le r\le\min\{d_x,d_y\}$.) Assume $\sigma_1\ge\cdots\ge\sigma_r$ W.L.O.G., so that $\|\boldsymbol{W}\|\le\ell$ is equivalent to $\sigma_1\le\ell$. Then we have

$$\begin{aligned}p^2\boldsymbol{x}^\top\boldsymbol{X}\boldsymbol{x} + 2pq\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{y} - q^2\boldsymbol{y}^\top\boldsymbol{Y}\boldsymbol{y} &= p^2\boldsymbol{x}^\top\boldsymbol{X}\boldsymbol{x} + 2pq\sum_{k=1}^r \sigma_k\boldsymbol{x}^\top\boldsymbol{u}_k\boldsymbol{v}_k^\top\boldsymbol{y} - q^2\boldsymbol{y}^\top\boldsymbol{Y}\boldsymbol{y} \\ &= p^2\boldsymbol{x}^\top\boldsymbol{X}\boldsymbol{x} + 2pq\sum_{k=1}^r \sigma_k\boldsymbol{u}_k^\top\boldsymbol{x}\boldsymbol{y}^\top\boldsymbol{v}_k^\top - q^2\boldsymbol{y}^\top\boldsymbol{Y}\boldsymbol{y}. \qquad (30)\end{aligned}$$

Since we aim to show an upper bound of (30), we now consider another optimization problem over a "bigger" search space and try to characterize its optimum value; this value will give us an upper bound of $\lambda_{\max}^{\boldsymbol{M}'}$. Namely, we now additionally treat $\boldsymbol{u}_1,\dots,\boldsymbol{u}_r$ and $\boldsymbol{v}_1,\dots,\boldsymbol{v}_r$ in (30) as optimization variables. With this addition, from now we treat the following items as optimization variables:

1. Choice of unit vectors $\boldsymbol{u}_1,\dots,\boldsymbol{u}_r\in\mathbb{R}^{d_x}$ of $\boldsymbol{U}$ and $\boldsymbol{v}_1,\dots,\boldsymbol{v}_r\in\mathbb{R}^{d_y}$ of $\boldsymbol{V}$
2. Choice of unit vectors $\boldsymbol{x}\in\mathbb{R}^{d_x}$, $\boldsymbol{y}\in\mathbb{R}^{d_y}$
3. Choice of values $p,q\in[0,1]$ such that $p^2+q^2 = 1$

Our problem boils down to finding the maximum value of (30) over all possible choices of these variables. (Note that the subsequent arguments and the resulting upper bound are true for all cases of $r\le\min\{d_x,d_y\}$.)

First, note that our choices of $\boldsymbol{u}_1,\dots,\boldsymbol{u}_r$ and $\boldsymbol{v}_1,\dots,\boldsymbol{v}_r$ only affect the middle term, which is bounded by

$$2pq\sum_{k=1}^r \sigma_k\boldsymbol{u}_k^\top\boldsymbol{x}\boldsymbol{y}^\top\boldsymbol{v}_k^\top \le 2pq\sigma_1,$$

for which, for any given $\boldsymbol{x}$, $\boldsymbol{y}$ and p,q, the maximum is attained when we choose $\boldsymbol{u}_1 = \boldsymbol{x}$, $\boldsymbol{v}_1 = \boldsymbol{y}$. (Note that the terms for $k\ge 2$ all disappear by orthogonality.)

Now we can observe that over possible choices of $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$p^2\boldsymbol{x}^\top\boldsymbol{X}\boldsymbol{x} + 2pq\sigma_1 - q^2\boldsymbol{y}^\top\boldsymbol{Y}\boldsymbol{y} \le p^2\lambda_{\max}^{\boldsymbol{X}} + 2pq\sigma_1 - q^2\lambda_{\min}^{\boldsymbol{Y}},$$

where equality holds if the unit vector $\boldsymbol{x}$ (or $\boldsymbol{y}$) is aligned with the eigenvector corresponding to the maximum (or minimum) eigenvalue of $\boldsymbol{X}$ (or $\boldsymbol{Y}$). We can use the given conditions to obtain

$$p^2\lambda_{\max}^{\boldsymbol{X}} + 2pq\sigma_1 - q^2\lambda_{\min}^{\boldsymbol{Y}} \le p^2 s_x + 2pq\ell - q^2 t_y = \begin{bmatrix}p\\ q\end{bmatrix}^\top\begin{bmatrix}s_x & \ell\\ \ell & -t_y\end{bmatrix}\begin{bmatrix}p\\ q\end{bmatrix}.$$

Finally, if we take the maximum over $p, q \in [0,1]$ with $p^2+q^2=1$, we have that

$$\sup_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \begin{bmatrix} p \\ q \end{bmatrix}^\top \begin{bmatrix} s_x & \ell \\ \ell & -t_y \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \left\| \begin{bmatrix} s_x & \ell \\ \ell & -t_y \end{bmatrix} \right\| = \left\| \begin{bmatrix} s_x & -\ell \\ \ell & t_y \end{bmatrix} \right\|$$

and hence we can conclude that

$$\lambda_{\max}^{\boldsymbol{M}'} \le \left\| \begin{bmatrix} s_x & -\ell \\ \ell & t_y \end{bmatrix} \right\|.$$

For the degenerate case $\ell = 0$, we can just apply $r = 1$ and $\sigma_1 = 0$, which does not hurt the validity of the proof.

Minimum Eigenvalue. Similarly, the minimum eigenvalue of $\boldsymbol{M}'$ is equal to

$$\begin{aligned}
\lambda_{\min}^{\boldsymbol{M}'} &= \inf_{\boldsymbol{z}\in\mathbb{R}^{d_x+d_y}, \|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \boldsymbol{M}' \boldsymbol{z} \\
&= \inf_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \inf_{\substack{\boldsymbol{x}\in\mathbb{R}^{d_x}, \|\boldsymbol{x}\|=1\\ \boldsymbol{y}\in\mathbb{R}^{d_y}, \|\boldsymbol{y}\|=1}} \begin{bmatrix} p\boldsymbol{x} \\ q\boldsymbol{y} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{X} & \boldsymbol{W} \\ \boldsymbol{W}^\top & -\boldsymbol{Y} \end{bmatrix} \begin{bmatrix} p\boldsymbol{x} \\ q\boldsymbol{y} \end{bmatrix} \\
&= \inf_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \inf_{\substack{\boldsymbol{x}\in\mathbb{R}^{d_x}, \|\boldsymbol{x}\|=1\\ \boldsymbol{y}\in\mathbb{R}^{d_y}, \|\boldsymbol{y}\|=1}} \left(p^2 \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} + 2pq \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{y} - q^2 \boldsymbol{y}^\top \boldsymbol{Y} \boldsymbol{y}\right) \\
&= - \sup_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^{d_x}, \|\boldsymbol{x}\|=1\\ \boldsymbol{y}\in\mathbb{R}^{d_y}, \|\boldsymbol{y}\|=1}} \left(-p^2 \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} - 2pq \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{y} + q^2 \boldsymbol{y}^\top \boldsymbol{Y} \boldsymbol{y}\right),
\end{aligned}$$

and therefore

$$-\lambda_{\min}^{\boldsymbol{M}'} = \sup_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^{d_x}, \|\boldsymbol{x}\|=1\\ \boldsymbol{y}\in\mathbb{R}^{d_y}, \|\boldsymbol{y}\|=1}} \left(-p^2 \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} - 2pq \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{y} + q^2 \boldsymbol{y}^\top \boldsymbol{Y} \boldsymbol{y}\right),$$

where we use the same reparameterization: $\boldsymbol{z} = \begin{bmatrix} p\boldsymbol{x}^\top & q\boldsymbol{y}^\top \end{bmatrix}^\top$ with $\boldsymbol{x} \in \mathbb{R}^{d_x}$, $\boldsymbol{y} \in \mathbb{R}^{d_y}$ with $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$, and $p^2 + q^2 = 1$.

As in the maximum case, we first assume that $\ell > 0$ and define the singular value decomposition of $\boldsymbol{W}$ as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. Then we can write

$$-p^2 \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} - 2pq \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{y} + q^2 \boldsymbol{y}^\top \boldsymbol{Y} \boldsymbol{y} = -p^2 \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} - 2pq \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k^\top \boldsymbol{x} \boldsymbol{y}^\top \boldsymbol{v}_k^\top + q^2 \boldsymbol{y}^\top \boldsymbol{Y} \boldsymbol{y}. \tag{31}$$

to observe that

$$-2pq \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k^\top \boldsymbol{x} \boldsymbol{y}^\top \boldsymbol{v}_k^\top \le 2pq\sigma_1,$$

for which the maximum is attained when we choose $\boldsymbol{u}_1 = \boldsymbol{x}$ and $\boldsymbol{v}_1 = -\boldsymbol{y}$. Then we have

$$-p^2 \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} + 2pq\sigma_1 + q^2 \boldsymbol{y}^\top \boldsymbol{Y} \boldsymbol{y} \le -p^2 \lambda_{\min}^{\boldsymbol{X}} + 2pq\sigma_1 + q^2 \lambda_{\max}^{\boldsymbol{Y}},$$

where equality holds if the unit vector $\boldsymbol{x}$ (or $\boldsymbol{y}$) is aligned with the eigenvector corresponding to the minimum (or maximum) eigenvalue of $\boldsymbol{X}$ (or $\boldsymbol{Y}$). We can use the given conditions to obtain

$$-p^2 \lambda_{\min}^{\boldsymbol{X}} + 2pq\sigma_1 + q^2 \lambda_{\max}^{\boldsymbol{Y}} \le -p^2 t_x + 2pq\ell - q^2 s_y = \begin{bmatrix} p \\ q \end{bmatrix}^\top \begin{bmatrix} t_x & \ell \\ \ell & -s_y \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix}.$$

Finally, if we take the maximum over $p, q \in [0,1]$ with $p^2+q^2=1$, we have that

$$\sup_{\substack{p,q\in[0,1]\\ p^2+q^2=1}} \begin{bmatrix} p \\ q \end{bmatrix}^\top \begin{bmatrix} t_x & \ell \\ \ell & -s_y \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \left\| \begin{bmatrix} t_x & \ell \\ \ell & -s_y \end{bmatrix} \right\| = \left\| \begin{bmatrix} t_x & -\ell \\ \ell & s_y \end{bmatrix} \right\|$$

and hence we can conclude that

$$-\lambda_{\min}^{\boldsymbol{M}'} \le \left\| \begin{bmatrix} t_x & -\ell \\ \ell & s_y \end{bmatrix} \right\|.$$

Combining the results with (26), we have

$$\|\boldsymbol{M}'\| = \max\left\{ \lambda_{\max}^{\boldsymbol{M}'}, -\lambda_{\min}^{\boldsymbol{M}'} \right\} = \max\left\{ \left\| \begin{bmatrix} s_x & -\ell \\ \ell & t_y \end{bmatrix} \right\|, \left\| \begin{bmatrix} t_x & -\ell \\ \ell & s_y \end{bmatrix} \right\| \right\}.$$

For the degenerate case $\ell = 0$, we can just apply $r = 1$ and $\sigma_1 = 0$, which does not hurt the validity of the proof.

Therefore we have shown (29), which completes the proof of Lemma B.5. □

Remark. An anonymous reviewer has found a much simpler proof of Lemma B.5. By definition we have

$$\begin{bmatrix} \|\boldsymbol{x}\| \\ \|\boldsymbol{y}\| \end{bmatrix}^\top \begin{bmatrix} t_x & -l \\ -l & -s_y \end{bmatrix} \begin{bmatrix} \|\boldsymbol{x}\| \\ \|\boldsymbol{y}\| \end{bmatrix} \le \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}^\top \boldsymbol{M}' \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \le \begin{bmatrix} \|\boldsymbol{x}\| \\ \|\boldsymbol{y}\| \end{bmatrix}^\top \begin{bmatrix} s_x & l \\ l & -t_y \end{bmatrix} \begin{bmatrix} \|\boldsymbol{x}\| \\ \|\boldsymbol{y}\| \end{bmatrix},$$

for all $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$. Then we immediately obtain the desired inequality as the matrix norm is invariant with respect to multiplication by -1 on rows and columns. □

B.4.4. Proof of Lemma B.6

Here we prove Lemma B.6, restated below for the sake of readability.

Lemma B.6. *Suppose that $A, B, C \ge 0$ and $A < B$. Then for the function $f : (0,\infty) \to (0,\infty)$ of the following form:*

$$f(x) = Ax + \sqrt{(1-Bx)^2 + C^2x^2},$$

the minimizer is equal to

$$x_\star = \frac{1}{D} \cdot \frac{2(C+D)(B-A)}{(C+D)^2 + (B-A)^2},$$

and the minimum value attained at $x_\star$ is equal to

$$f(x_\star) = \frac{(C+D)^2 - (B-A)^2}{(C+D)^2 + (B-A)^2},$$

where $D = \sqrt{B^2 + C^2 - A^2}$.

Proof. Observing that $A^2 + D^2 = B^2 + C^2$ by definition, we start by substituting

$$R = \sqrt{B^2+C^2} = \sqrt{A^2+D^2}, \;\; \sin\phi = \frac{A}{R}, \;\; \sin\psi = \frac{B}{R}$$

for $\phi \in [0, \frac{\pi}{2})$ and $\psi \in (0, \frac{\pi}{2}]$. Note that we have $\phi < \psi$ from $A < B$, and

$$\cos\phi = \frac{D}{R}, \;\; \cos\psi = \frac{C}{R}.$$

We can compute

$$\begin{aligned} Ax + \sqrt{(1-Bx)^2 + C^2x^2} &= Rx\sin\phi + \sqrt{(1 - Rx\sin\psi)^2 + R^2x^2\cos^2\psi} \\ &= Rx\sin\phi + \sqrt{1 - 2Rx\sin\psi + R^2x^2}. \end{aligned}$$

By using change of variables as

$$y = \tan\psi - Rx\sec\psi \;\Leftrightarrow\; x = \frac{1}{R}\left(\sin\psi - y\cos\psi\right),$$

we have $y \in [-\infty, \tan\psi]$, and

$$1 - 2xR\sin\psi + R^2x^2 = (1+y^2)\cos^2\psi.$$

Plugging in, we can obtain the following reparameterization:

$$Rx\sin\phi + \sqrt{1 - 2Rx\sin\psi + R^2x^2} = \sin\phi\sin\psi - y\sin\phi\cos\psi + \sqrt{1+y^2}\cos\psi.$$

We can easily observe that if we again reparameterize as $y = \sinh\theta$, we can write as

$$\sin\phi\sin\psi - \sin\phi\cos\psi\cdot\sinh\theta + \cos\psi\cdot\cosh\theta = \sin\phi\sin\psi + \cos\psi\left(\cosh\theta - \sin\phi\cdot\sinh\theta\right). \tag{32}$$

The derivative of (32) with respect to θ is equal to

$$\cos\psi\left(\sinh\theta - \sin\phi\cdot\cosh\theta\right). \tag{33}$$

As the *second* derivative of (32) satisfies $\cos\psi\left(\cosh\theta - \sin\phi\cdot\sinh\theta\right) \geq \cos\psi\cdot\left(-\sinh\theta + \cosh\theta\right) \geq 0$, we have that (33) is an increasing function. Therefore, the minimizer of (32) must be equal to the point where (33) is zero, which is[7]

$$y_\star = \sinh\theta_\star = \frac{\sin\phi}{\sqrt{1-\sin^2\phi}} = \frac{\sin\phi}{\cos\phi} = \tan\phi.$$

Note that we have $\cos\phi > 0$ since $\phi \in [0, \frac{2}{\pi})$, and using the square root expression above, we can compute

$$\cosh\theta_\star - \sin\phi\cdot\sinh\theta_\star = \frac{1}{\sqrt{1-\sin^2\phi}} - \frac{\sin^2\phi}{\sqrt{1-\sin^2\phi}} = \sqrt{1-\sin^2\phi} = \cos\phi. \tag{34}$$

The range of y contains $y_\star$, since $\phi < \psi$ implies $\tan\phi \in [-\infty, \tan\psi)$. We can substitute back as

$$x_\star = \frac{1}{R}\left(\sin\psi - \tan\phi\cos\psi\right) = \frac{1}{R\cos\phi}\left(\cos\phi\sin\psi - \sin\phi\cos\psi\right) = \frac{1}{R\cos\phi}\sin(\psi-\phi).$$

By using the trigonometric identity:

$$\frac{\sin\psi - \sin\phi}{\cos\psi + \cos\phi} = \frac{2\cos\left(\frac{\psi+\phi}{2}\right)\sin\left(\frac{\psi-\phi}{2}\right)}{2\cos\left(\frac{\psi+\phi}{2}\right)\cos\left(\frac{\psi-\phi}{2}\right)} = \tan\left(\frac{\psi-\phi}{2}\right), \tag{35}$$

we can compute

$$\sin(\psi-\phi) = \frac{2\tan\left(\frac{\psi-\phi}{2}\right)}{1+\tan^2\left(\frac{\psi-\phi}{2}\right)} = \frac{2(\cos\psi+\cos\phi)(\sin\psi-\sin\phi)}{(\cos\psi+\cos\phi)^2 + (\sin\psi-\sin\phi)^2},$$

and combined with $D = R\cos\phi$ we can conclude that

$$x_\star = \frac{1}{D}\cdot\frac{2(C+D)(B-A)}{(C+D)^2+(B-A)^2}.$$

Also, by (34), the minimum value can also be computed as

$$f_\star = \sin\phi\sin\psi + \cos\psi\left(\cosh\theta_\star - \sin\phi\cdot\sinh\theta_\star\right) = \sin\phi\sin\psi + \cos\phi\cos\psi = \cos(\psi-\phi).$$

[7] To clarify, we are just using the fact that $a\sinh t - b\cosh t = 0$ if $\sinh t = \frac{b}{\sqrt{a^2-b_2}}$.

By using the trigonometric identity in (35), we can compute

$$\cos(\psi-\phi)=\frac{1-\tan^2\left(\frac{\psi-\phi}{2}\right)}{1+\tan^2\left(\frac{\psi-\phi}{2}\right)}=\frac{(\cos\psi+\cos\phi)^2-(\sin\psi-\sin\phi)^2}{(\cos\psi+\cos\phi)^2+(\sin\psi-\sin\phi)^2},$$

and we can conclude that

$$f_\star=\frac{(\cos\psi+\cos\phi)^2-(\sin\psi-\sin\phi)^2}{(\cos\psi+\cos\phi)^2+(\sin\psi-\sin\phi)^2}=\frac{(C+D)^2-(B-A)^2}{(C+D)^2+(B-A)^2}$$

as desired. □

B.4.5. Proof of Lemma B.7

Here we prove Lemma B.7, restated below for the sake of readability.

Lemma B.7. *Suppose that $A_1, A_2, B_1, B_2, C \geq 0$ satisfies $A_1 \leq B_1$, $A_2 \leq B_2$, and $A_2 - A_1 = B_2 - B_1 \geq 0$. Then for the functions $f_1, f_2 : (0,\infty) \to (0,\infty)$ of the following form:*

$$f_1(x)=A_1x+\sqrt{(1-B_1x)^2+C^2x^2},\quad f_2(x)=A_2x+\sqrt{(1-B_2x)^2+C^2x^2},$$

we have $f_1(x) \leq f_2(x)$ for all $x > 0$.

Proof. We must show that for all $x > 0$ we have $f_1(x) \leq f_2(x)$, i.e.,

$$A_1x+\sqrt{(1-B_1x)^2+C^2x^2}\leq A_2x+\sqrt{(1-B_2x)^2+C^2x^2},$$

which is equivalent to

$$\sqrt{\left(\frac{1}{x}-B_1\right)^2+C^2}-\sqrt{\left(\frac{1}{x}-B_2\right)^2+C^2}\leq A_2-A_1.$$

Let us substitute as follows:

$$D=A_2-A_1=B_2-B_1,\quad s=\frac{1}{x}-\frac{B_1+B_2}{2},$$

where $D \geq 0$ and $s > -\frac{B_1+B_2}{2}$ by assumption. We are left to show that

$$\sqrt{\left(s+\frac{D}{2}\right)^2+C^2}-\sqrt{\left(s-\frac{D}{2}\right)^2+C^2}\leq D. \tag{36}$$

If $D = 0$, then we can observe that both sides become 0 and hence (36) is indeed true.

If $D > 0$, the LHS of (36) as a function of s is a (monotonically) increasing function. Moreover, since

$$\lim_{s\to-\infty}\sqrt{\left(s+\frac{D}{2}\right)^2+C^2}-\sqrt{\left(s-\frac{D}{2}\right)^2+C^2}=-D,$$
$$\lim_{s\to\infty}\sqrt{\left(s+\frac{D}{2}\right)^2+C^2}-\sqrt{\left(s-\frac{D}{2}\right)^2+C^2}=D,$$

the range of the LHS is equal to $(-D, D)$, including when $C = 0$, which completes the proof. □

C. Proofs used in Section 4

Here we prove all theorems related to **Alt-GDA** presented in Section 4.

- In Appendix C.1 we prove Theorem 4.1 which yields a contraction inequality for **Alt-GDA**.
- In Appendix C.2 we prove Corollary 4.2 which derives the corresponding iteration complexity upper bound.
- In Appendix C.3 we prove the two main propositions introduced in Appendix C.

Notations. For notational simplicity, in Appendix B we define and use the following notations for gradients:

$$\boldsymbol{g}_{ij}^{x} := \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_j), \quad \boldsymbol{g}_{ij}^{y} := \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_i, \boldsymbol{y}_j).$$

In particular, we will use indices $i, j \in \{0, 1, \star\}$ throughout the proof.

C.1. Proof of Theorem 4.1

Here we prove Theorem 4.1 of Section 4, restated below for the sake of readability.

Theorem 4.1. *Suppose that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and we run **Alt-GDA** with step sizes $\alpha, \beta > 0$ that satisfy*

$$\alpha \leq \frac{1}{2} \cdot \min\left\{\frac{1}{L_x}, \frac{\sqrt{\mu_y}}{L_{xy}\sqrt{L_x}}\right\},$$
$$\beta \leq \frac{1}{2} \cdot \min\left\{\frac{1}{L_y}, \frac{\sqrt{\mu_x}}{L_{xy}\sqrt{L_y}}\right\}.$$

Then Ψ_k^{Alt} is valid, and satisfies $\Psi_{k+1}^{\text{Alt}} \leq r\Psi_k^{\text{Alt}}$ with

$$r = \max\left\{\frac{\frac{1}{\alpha} - \mu_x}{\frac{1}{\alpha} - 2\beta^2 L_y L_{xy}^2}, \frac{\frac{1}{\beta} - \mu_y}{\frac{1}{\beta} - \alpha^2 L_x L_{xy}^2}, \frac{\frac{1}{\alpha} - \mu_x}{\frac{1}{\alpha}}\right\},$$

where we have $0 < r < 1$.

Proof. Note that the Lyapunov function Ψ_k^{Alt} for **Alt-GDA** can be written as

$$\begin{aligned}\Psi_k^{\text{Alt}} = &\left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_k - \boldsymbol{y}_\star\|^2 \\ &+ \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|^2 - \alpha(1 - \alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2.\end{aligned} \tag{37}$$

The proof consists of two steps; in STEP 1 we prove that Ψ_k^{Alt} is a valid Lyapunov function, and in STEP 2 we show that $\Psi_{k+1}^{\text{Alt}} \leq r\Psi_k^{\text{Alt}}$ holds for the contraction rate r given as in Theorem 4.1. For notational simplicity, W.L.O.G. we equivalently show that the statement holds for $k = 0$ and *any* choice of initialization $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. (This is indeed safe because we can apply the results to each of the iterates of the whole sequence $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k\geq 0}$ generated by **Alt-GDA**.)

STEP 1. VALIDITY OF LYAPUNOV FUNCTION

Here we show that there exists some constant A^{Alt} such that we have $\Psi_0^{\text{Alt}} \geq A^{\text{Alt}}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2\right)$ for any choice of initialization $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, which is equivalent to showing that Ψ_k^{Alt} is a valid Lyapunov function. Proposition C.1 yields a lower bound inequality from which we can derive such a constant A^{Alt}.

Proposition C.1. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and **Alt-GDA** with step sizes given as in Theorem 4.1, we have*

$$\Psi_0^{\text{Alt}} \geq \left(\frac{1}{2\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{3}{4\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2. \tag{38}$$

for any choice of initialization $(\boldsymbol{x}_0, \boldsymbol{y}_0)$.

While we defer the proof of Proposition C.1 to Appendix C.3.1, here we see that

$$\left(\frac{1}{2\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+2\left(\frac{3}{4\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2\geq A^{\mathrm{Alt}}\left(\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2\right)$$

shows the validity of Ψ_0^{Alt} for $A^{\mathrm{Alt}}=\min\left\{\frac{1}{2\alpha}-\mu_x, 2\left(\frac{3}{4\beta}-\mu_y\right)\right\}>0$.

(Note that $\alpha\leq\frac{1}{2L_x}<\frac{1}{2\mu_x}$ and $\beta\leq\frac{1}{2L_y}<\frac{1}{2\mu_y}<\frac{3}{4\mu_y}$ implies $A^{\mathrm{Alt}}>0$.)

Step 2. Contraction Inequality

Here we show that $\Psi_1^{\mathrm{Alt}}\leq r\Psi_0^{\mathrm{Alt}}$ for any choice of initialization $(\boldsymbol{x}_0,\boldsymbol{y}_0)$, which is equivalent to showing that $\Psi_{k+1}^{\mathrm{Alt}}\leq r\Psi_k^{\mathrm{Alt}}$ for all k. Proposition C.2 yields a one-step contraction inequality that applies to **Alt-GDA** with $\alpha<\frac{1}{2L_x}$ and $\beta<\frac{1}{2L_y}$, *i.e.*, when the step sizes are small enough.

Proposition C.2. *For $f\in\mathcal{F}(\mu_x,\mu_y,L_x,L_y,L_{xy})$ and **Alt-GDA** with step sizes $\alpha\leq\frac{1}{2L_x}$ and $\beta\leq\frac{1}{2L_y}$, we have*

$$\begin{aligned}&\left(\frac{1}{\alpha}-2\beta^2L_yL_{xy}^2\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+2\left(\frac{1}{\beta}-\alpha^2L_xL_{xy}^2\right)\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2+\frac{1}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2-\alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^x\|^2\\&\leq\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2-\alpha(1-\alpha L_x)\|\boldsymbol{g}_{00}^x\|^2\end{aligned}\tag{39}$$

for all $\boldsymbol{x}_0\in\mathbb{R}^{d_x},\boldsymbol{y}_0\in\mathbb{R}^{d_y}$.

We prove Proposition C.2 in Appendix C.3.2.

Note that the choices of step sizes in Theorem 4.1 indeed satisfy $\alpha\leq\frac{1}{2L_x}$ and $\beta\leq\frac{1}{2L_y}$. Assume W.L.O.G. that $\boldsymbol{x}_\star=\mathbf{0}$ $(\in\mathbb{R}^{d_x})$ and $\boldsymbol{y}_\star=\mathbf{0}$ $(\in\mathbb{R}^{d_y})$. Observing that the RHS of (39) is exactly Ψ_0^{Alt}, it is enough to show that

$$\begin{aligned}\Psi_1^{\mathrm{Alt}}&=\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2+\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2-\alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^x\|^2\\&\leq r\left(\frac{1}{\alpha}-2\beta^2L_yL_{xy}^2\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+2r\left(\frac{1}{\beta}-\alpha^2L_xL_{xy}^2\right)\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2+\frac{r}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2-r\alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^x\|^2,\end{aligned}\tag{40}$$

after which we can combine the results as $r\cdot(39)+(40)$ to obtain $\Psi_1^{\mathrm{Alt}}\leq r\Psi_0^{\mathrm{Alt}}$.

Since $r\geq\frac{\frac{1}{\alpha}-\mu_x}{\frac{1}{\alpha}-2\beta^2L_yL_{xy}^2}$, we have

$$\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2\leq r\left(\frac{1}{\alpha}-2\beta^2L_yL_{xy}^2\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2.$$

Since $r\geq\frac{\frac{1}{\beta}-\mu_y}{\frac{1}{\beta}-\alpha^2L_xL_{xy}^2}$, we have

$$2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2\leq 2r\left(\frac{1}{\beta}-\alpha^2L_xL_{xy}^2\right)\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2.$$

Since $r\geq\frac{\frac{1}{\alpha}-\mu_x}{\frac{1}{\alpha}}$, we have

$$\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2\leq\frac{r}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2.$$

Also, we can observe that $\alpha\leq\frac{1}{2L_x}$ and $\beta\leq\frac{1}{2}\sqrt{\frac{\mu_x}{L_y}}\cdot\frac{1}{L_{xy}}$ implies

$$\alpha\beta^2\leq\frac{1}{2L_x}\cdot\frac{\mu_x}{4L_yL_{xy}^2}<\frac{1}{2L_yL_{xy}^2}\quad\wedge\quad 2\beta^2L_yL_{xy}^2<4\beta^2L_yL_{xy}^2\leq\mu_x\quad\Rightarrow\quad\frac{\frac{1}{\alpha}-\mu_x}{\frac{1}{\alpha}-2\beta^2L_yL_{xy}^2}\in(0,1),$$

and that $\alpha \leq \frac{1}{2}\sqrt{\frac{\mu_y}{L_x}} \cdot \frac{1}{L_{xy}}$ and $\beta \leq \frac{1}{2L_y}$ implies

$$\alpha^2\beta \leq \frac{\mu_y}{4L_xL_{xy}^2} \cdot \frac{1}{2L_y} < \frac{1}{L_xL_{xy}^2} \quad \wedge \quad \alpha^2 L_x L_{xy}^2 < 4\alpha^2 L_x L_{xy}^2 \leq \mu_y \quad \Rightarrow \quad \frac{\frac{1}{\beta} - \mu_y}{\frac{1}{\beta} - \alpha^2 L_x L_{xy}^2} \in (0,1).$$

Since it is obvious that $\frac{\frac{1}{\alpha}-\mu_x}{\frac{1}{\alpha}} \in (0,1)$, we can observe that

$$r = \max\left\{\frac{\frac{1}{\alpha}-\mu_x}{\frac{1}{\alpha}-2\beta^2 L_y L_{xy}^2}, \frac{\frac{1}{\beta}-\mu_y}{\frac{1}{\beta}-\alpha^2 L_x L_{xy}^2}, \frac{\frac{1}{\alpha}-\mu_x}{\frac{1}{\alpha}}\right\} \in (0,1)$$

and therefore

$$-\alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^x\|^2 \leq -r\alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^x\|^2,$$

which shows $r \in (0,1)$ and (40), and–altogether with Proposition C.2–proves the given statement. □

C.2. Proof of Corollary 4.2

Here we prove Corollary 4.2 of Section 4, restated below for the sake of readability.

Corollary 4.2. *For step sizes given by the maximum possible values in Theorem 4.1, **Alt-GDA** linearly converges with iteration complexity*

$$\mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}(\sqrt{\kappa_x} + \sqrt{\kappa_y})\right) \cdot \log \frac{\Psi_0^{\text{Alt}}}{A^{\text{Alt}}\epsilon}\right),$$

where $A^{\text{Alt}} = \min\left\{\frac{1}{2\alpha} - \mu_x, 2\left(\frac{3}{4\beta} - \mu_y\right)\right\} > 0.$

Proof. From Theorem 4.1, we have

$$\frac{1}{1-r} = \max\left\{\frac{\frac{1}{\alpha}-2\beta^2 L_y L_{xy}^2}{\mu_x - 2\beta^2 L_y L_{xy}^2}, \frac{\frac{1}{\beta}-\alpha^2 L_x L_{xy}^2}{\mu_y - \alpha^2 L_x L_{xy}^2}, \frac{1}{\alpha\mu_x}\right\}.$$

From $\beta \leq \frac{1}{2} \cdot \sqrt{\frac{\mu_x}{L_y}} \cdot \frac{1}{L_{xy}}$, we have

$$\frac{\frac{1}{\alpha}-2\beta^2 L_y L_{xy}^2}{\mu_x - 2\beta^2 L_y L_{xy}^2} \leq \frac{\frac{1}{\alpha}-\frac{1}{2}\mu_x}{\mu_x - \frac{1}{2}\mu_x} \leq \frac{2}{\alpha\mu_x}.$$

From $\alpha \leq \frac{1}{2} \cdot \sqrt{\frac{\mu_y}{L_x}} \cdot \frac{1}{L_{xy}}$, we have

$$\frac{\frac{1}{\beta}-\alpha^2 L_x L_{xy}^2}{\mu_y - \alpha^2 L_x L_{xy}^2} \leq \frac{\frac{1}{\beta}-\frac{1}{4}\mu_y}{\mu_y - \frac{1}{4}\mu_y} \leq \frac{4}{3\beta\mu_y}.$$

We can deduce that

$$\begin{aligned}\frac{1}{1-r} \leq \max\left\{\frac{2}{\alpha\mu_x}, \frac{4}{3\beta\mu_y}\right\} &= \max\left\{\Theta\left(\kappa_x + \kappa_{xy}\sqrt{\kappa_x}\right), \Theta\left(\kappa_y + \kappa_{xy}\sqrt{\kappa_y}\right)\right\} \\ &= \Theta\left(\kappa_x + \kappa_y + \kappa_{xy}(\sqrt{\kappa_x} + \sqrt{\kappa_y})\right).\end{aligned}$$

Therefore it is sufficient to take

$$K = \mathcal{O}\left(\left(\kappa_x + \kappa_y + \kappa_{xy}(\sqrt{\kappa_x} + \sqrt{\kappa_y})\right) \cdot \log \frac{\Psi_0^{\text{Alt}}}{A^{\text{Alt}}\epsilon}\right)$$

iterations to ensure that $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$.

Finally, we can check that $\alpha \leq \frac{1}{2L_x} < \frac{1}{2\mu_x}$ and $\beta \leq \frac{1}{2L_y} < \frac{1}{2\mu_y} < \frac{3}{4\mu_y}$ implies $A^{\text{Alt}} > 0$. □

C.3. Proofs used in Appendix C

Here we prove the propositions introduced in Appendix C.

C.3.1. Proof of Proposition C.1

Here we prove Proposition C.1, restated below for the sake of readability.

Proposition C.1. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and **Alt-GDA** with step sizes given as in Theorem 4.1, we have*

$$\Psi_0^{\text{Alt}} \geq \left(\frac{1}{2\alpha} - \mu_x\right) \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{3}{4\beta} - \mu_y\right) \|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2. \tag{38}$$

for any choice of initialization $(\boldsymbol{x}_0, \boldsymbol{y}_0)$.

Proof. For simplicity let us assume W.L.O.G. that $\boldsymbol{x}_\star = \mathbf{0}$ $(\in \mathbb{R}^{d_x})$ and $\boldsymbol{y}_\star = \mathbf{0}$ $(\in \mathbb{R}^{d_y})$.

By triangle inequality and Lipschitz gradients, we have

$$\|\boldsymbol{g}_{00}^x\|^2 \leq 2\|\boldsymbol{g}_{00}^x - \boldsymbol{g}_{\star 0}^x\|^2 + 2\|\boldsymbol{g}_{\star 0}^x\|^2 \leq 2L_x^2\|\boldsymbol{x}_0\|^2 + 2L_{xy}^2\|\boldsymbol{y}_0\|^2.$$

Therefore, we can obtain

$$\begin{aligned}
&\left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right) \|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_1\|^2 - \alpha(1 - \alpha L_x)\|\boldsymbol{g}_{00}^x\|^2 \\
&\geq \left(\frac{1}{\alpha} - \mu_x - 2\alpha(1 - \alpha L_x)L_x^2\right) \|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta} - \mu_y - \alpha(1 - \alpha L_x)L_{xy}^2\right) \|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_1\|^2.
\end{aligned}$$

Since $\alpha \leq \frac{1}{2L_x}$, we have

$$\frac{1}{\alpha} - \mu_x - 2\alpha(1 - \alpha L_x)L_x^2 \geq \frac{1}{\alpha} - \mu_x - 2\alpha L_x^2 \geq \frac{1}{\alpha} - \mu_x - \frac{1}{2\alpha} = \frac{1}{2\alpha} - \mu_x.$$

Since $\alpha \leq \frac{1}{2}\sqrt{\frac{\mu_y}{L_x}} \cdot \frac{1}{L_{xy}}$ and $\beta \leq \frac{1}{2}\sqrt{\frac{\mu_x}{L_y}} \cdot \frac{1}{L_{xy}}$, we have

$$\frac{1}{\beta} - \mu_y - \alpha(1 - \alpha L_x)L_{xy}^2 \geq \frac{1}{\beta} - \mu_y - \alpha L_{xy}^2 \geq \frac{1}{\beta} - \mu_y - \frac{1}{4\beta}\sqrt{\frac{\mu_y}{L_x}} \cdot \sqrt{\frac{\mu_x}{L_y}} \geq \frac{1}{\beta} - \mu_y - \frac{1}{4\beta} = \frac{3}{4\beta} - \mu_y.$$

Therefore we have

$$\begin{aligned}
&\left(\frac{1}{\alpha} - \mu_x - 2\alpha(1 - \alpha L_x)L_x^2\right) \|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta} - \mu_y - \alpha(1 - \alpha L_x)L_{xy}^2\right) \|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_1\|^2 \\
&\geq \left(\frac{1}{2\alpha} - \mu_x\right) \|\boldsymbol{x}_0\|^2 + 2\left(\frac{3}{4\beta} - \mu_y\right) \|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_1\|^2,
\end{aligned}$$

which proves that (38) is indeed true. □

C.3.2. Proof of Proposition C.2

Here we prove Proposition C.2, restated below for the sake of readability.

Proposition C.2. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and **Alt-GDA** with step sizes $\alpha \leq \frac{1}{2L_x}$ and $\beta \leq \frac{1}{2L_y}$, we have*

$$\begin{aligned}
&\left(\frac{1}{\alpha} - 2\beta^2 L_y L_{xy}^2\right) \|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \alpha^2 L_x L_{xy}^2\right) \|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 - \alpha(1 - \alpha L_x)\|\boldsymbol{g}_{11}^x\|^2 \\
&\leq \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right) \|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right) \|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 - \alpha(1 - \alpha L_x)\|\boldsymbol{g}_{00}^x\|^2
\end{aligned} \tag{39}$$

for all $\boldsymbol{x}_0 \in \mathbb{R}^{d_x}, \boldsymbol{y}_0 \in \mathbb{R}^{d_y}$.

Proof. Recall that **Alt-GDA** takes updates of the form:

$$\begin{aligned}\boldsymbol{x}_1 &= \boldsymbol{x}_0 - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0) = \boldsymbol{x}_0 - \alpha \boldsymbol{g}_{00}^x,\\ \boldsymbol{y}_1 &= \boldsymbol{y}_0 + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_1, \boldsymbol{y}_0) = \boldsymbol{y}_0 + \beta \boldsymbol{g}_{10}^y.\end{aligned}$$

From this, we can deduce that

$$\begin{aligned}\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 &= \frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \frac{2}{\alpha}\left\langle \boldsymbol{x}_1 - \boldsymbol{x}_0, \boldsymbol{x}_1 - \boldsymbol{x}_\star\right\rangle - \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2\\ &= \frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 - 2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_1 - \boldsymbol{x}_\star\right\rangle - \alpha\|\boldsymbol{g}_{00}^x\|^2,\\ \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 &= \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \frac{2}{\beta}\left\langle \boldsymbol{y}_1 - \boldsymbol{y}_0, (\boldsymbol{y}_0 - \boldsymbol{y}_\star) + (\boldsymbol{y}_1 - \boldsymbol{y}_\star)\right\rangle\\ &= \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + 2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_0 - \boldsymbol{y}_\star\right\rangle + 2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_1 - \boldsymbol{y}_\star\right\rangle,\\ \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 &= \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\alpha}\left\langle \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{x}_1 - \boldsymbol{x}_\star\right\rangle + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|^2\\ &= \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 - 2\left\langle \boldsymbol{g}_{11}^x, \boldsymbol{x}_1 - \boldsymbol{x}_\star\right\rangle + \alpha\|\boldsymbol{g}_{11}^x\|^2,\end{aligned}$$

which sums up to

$$\begin{aligned}&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2\\ &= \frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 - \alpha\|\boldsymbol{g}_{00}^x\|^2 + \alpha\|\boldsymbol{g}_{11}^x\|^2\\ &\quad - 2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_1 - \boldsymbol{x}_\star\right\rangle + 2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_0 - \boldsymbol{y}_\star\right\rangle + 2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_1 - \boldsymbol{y}_\star\right\rangle - 2\left\langle \boldsymbol{g}_{11}^x, \boldsymbol{x}_1 - \boldsymbol{x}_\star\right\rangle.\end{aligned} \tag{41}$$

Then μ_x-strong convexity and L_x-Lipschitz gradients[8] of $f(\cdot, \boldsymbol{y}_0)$ yields:

$$2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_\star - \boldsymbol{x}_0\right\rangle = 2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0), \boldsymbol{x}_\star - \boldsymbol{x}_0\right\rangle \le -\mu_x\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 - 2(f(\boldsymbol{x}_0, \boldsymbol{y}_0) - f(\boldsymbol{x}_\star, \boldsymbol{y}_0)), \tag{42}$$

$$2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_0 - \boldsymbol{x}_1\right\rangle = -2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0), \boldsymbol{x}_1 - \boldsymbol{x}_0\right\rangle \le L_x\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + 2(f(\boldsymbol{x}_0, \boldsymbol{y}_0) - f(\boldsymbol{x}_1, \boldsymbol{y}_0)). \tag{43}$$

Similarly, μ_y-strong concavity and L_y-Lipschitz gradients of $f(\boldsymbol{x}_1, \cdot)$ yields:

$$2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_0 - \boldsymbol{y}_\star\right\rangle = -2\left\langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_1, \boldsymbol{y}_0), \boldsymbol{y}_\star - \boldsymbol{y}_0\right\rangle \le -\mu_y\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 - 2(f(\boldsymbol{x}_1, \boldsymbol{y}_\star) - f(\boldsymbol{x}_1, \boldsymbol{y}_0)), \tag{44}$$

$$2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_1 - \boldsymbol{y}_0\right\rangle = 2\left\langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_1, \boldsymbol{y}_0), \boldsymbol{y}_1 - \boldsymbol{y}_0\right\rangle \le L_y\|\boldsymbol{y}_1 - \boldsymbol{y}_0\|^2 + 2(f(\boldsymbol{x}_1, \boldsymbol{y}_1) - f(\boldsymbol{x}_1, \boldsymbol{y}_0)). \tag{45}$$

Finally, μ_x-strong convexity of $f(\cdot, \boldsymbol{y}_1)$ yields:

$$2\left\langle \boldsymbol{g}_{11}^x, \boldsymbol{x}_\star - \boldsymbol{x}_1\right\rangle = 2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \boldsymbol{y}_1), \boldsymbol{x}_\star - \boldsymbol{x}_1\right\rangle \le -\mu_x\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 - 2(f(\boldsymbol{x}_1, \boldsymbol{y}_1) - f(\boldsymbol{x}_\star, \boldsymbol{y}_1)). \tag{46}$$

From now, for simplicity we assume W.L.O.G. $\boldsymbol{x}_\star = \boldsymbol{0}$ $(\in \mathbb{R}^{d_x})$ and $\boldsymbol{y}_\star = \boldsymbol{0}$ $(\in \mathbb{R}^{d_y})$.

From (42) + (43) we have

$$\begin{aligned}-2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_1\right\rangle &= 2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_\star - \boldsymbol{x}_0\right\rangle + 2\left\langle \boldsymbol{g}_{00}^x, \boldsymbol{x}_0 - \boldsymbol{x}_1\right\rangle\\ &\le -\mu_x\|\boldsymbol{x}_0\|^2 + L_x\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + 2(f(\boldsymbol{x}_\star, \boldsymbol{y}_0) - f(\boldsymbol{x}_1, \boldsymbol{y}_0))\\ &= -\mu_x\|\boldsymbol{x}_0\|^2 + \alpha^2 L_x\|\boldsymbol{g}_{00}^x\|^2 + 2(f(\boldsymbol{x}_\star, \boldsymbol{y}_0) - f(\boldsymbol{x}_1, \boldsymbol{y}_0)).\end{aligned}$$

From $2\times$ (44) + (45) we have

$$\begin{aligned}2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_0\right\rangle + 2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_1\right\rangle &= 4\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_0 - \boldsymbol{y}_\star\right\rangle + 2\left\langle \boldsymbol{g}_{10}^y, \boldsymbol{y}_1 - \boldsymbol{y}_0\right\rangle\\ &\le -2\mu_y\|\boldsymbol{y}_0\|^2 + L_y\|\boldsymbol{y}_1 - \boldsymbol{y}_0\|^2 - 2(2f(\boldsymbol{x}_1, \boldsymbol{y}_\star) - f(\boldsymbol{x}_1, \boldsymbol{y}_0) - f(\boldsymbol{x}_1, \boldsymbol{y}_1))\\ &= -2\mu_y\|\boldsymbol{y}_0\|^2 + \beta^2 L_y\|\boldsymbol{g}_{10}^y\|^2 - 2(2f(\boldsymbol{x}_1, \boldsymbol{y}_\star) - f(\boldsymbol{x}_1, \boldsymbol{y}_0) - f(\boldsymbol{x}_1, \boldsymbol{y}_1)).\end{aligned}$$

[8] Note that the Lipschitz gradient conditions for L_x and L_y are equivalent to the widely used notion of *smoothness* in convex optimization literature.

Finally, (46) translates into

$$-2\left\langle \boldsymbol{g}_{11}^{x}, \boldsymbol{x}_1\right\rangle = 2\left\langle \boldsymbol{g}_{11}^{x}, \boldsymbol{x}_\star - \boldsymbol{x}_1\right\rangle \le -\mu_x\|\boldsymbol{x}_1\|^2 - 2(f(\boldsymbol{x}_1,\boldsymbol{y}_1) - f(\boldsymbol{x}_\star,\boldsymbol{y}_1)).$$

We can properly plug in the above equations to Equation (41) to obtain

$$\begin{aligned}\frac{1}{\alpha}\|\boldsymbol{x}_1\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2\|^2 \le{}& \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1\|^2\\ &- \alpha(1-\alpha L_x)\|\boldsymbol{g}_{00}^{x}\|^2 + \alpha\|\boldsymbol{g}_{11}^{x}\|^2 + \beta^2 L_y\|\boldsymbol{g}_{10}^{y}\|^2\\ &- 2(2f(\boldsymbol{x}_1,\boldsymbol{y}_\star) - f(\boldsymbol{x}_\star,\boldsymbol{y}_0) - f(\boldsymbol{x}_\star,\boldsymbol{y}_1)).\end{aligned}$$

Since f is convex-concave and has Lipschitz gradients, we have

$$\begin{aligned}-2(f(\boldsymbol{x}_1,\boldsymbol{y}_\star) - f(\boldsymbol{x}_\star,\boldsymbol{y}_\star)) &\le -\frac{1}{L_x}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\boldsymbol{y}_\star)\|^2 = -\frac{1}{L_x}\|\boldsymbol{g}_{1\star}^{x}\|^2,\\ -2(f(\boldsymbol{x}_\star,\boldsymbol{y}_\star) - f(\boldsymbol{x}_\star,\boldsymbol{y}_0)) &\le -\frac{1}{L_y}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_0)\|^2 = -\frac{1}{L_y}\|\boldsymbol{g}_{\star 0}^{y}\|^2,\\ -2(f(\boldsymbol{x}_\star,\boldsymbol{y}_\star) - f(\boldsymbol{x}_\star,\boldsymbol{y}_1)) &\le -\frac{1}{L_y}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_\star,\boldsymbol{y}_1)\|^2 = -\frac{1}{L_y}\|\boldsymbol{g}_{\star 1}^{y}\|^2.\end{aligned}$$

Therefore we have

$$\begin{aligned}\frac{1}{\alpha}\|\boldsymbol{x}_1\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2\|^2 \le{}& \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1\|^2\\ &- \alpha(1-\alpha L_x)\|\boldsymbol{g}_{00}^{x}\|^2 + \alpha\|\boldsymbol{g}_{11}^{x}\|^2 - \frac{2}{L_x}\|\boldsymbol{g}_{1\star}^{x}\|^2\\ &+ \beta^2 L_y\|\boldsymbol{g}_{10}^{y}\|^2 - \frac{1}{L_y}\|\boldsymbol{g}_{\star 0}^{y}\|^2 - \frac{1}{L_y}\|\boldsymbol{g}_{\star 1}^{y}\|^2\\ ={}& \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1\|^2\\ &+ \alpha^2 L_x\|\boldsymbol{g}_{11}^{x}\|^2 - \frac{2}{L_x}\|\boldsymbol{g}_{1\star}^{x}\|^2 + \beta^2 L_y\|\boldsymbol{g}_{10}^{y}\|^2 - \frac{1}{L_y}\|\boldsymbol{g}_{\star 0}^{y}\|^2\\ &- \alpha(1-\alpha L_x)\|\boldsymbol{g}_{00}^{x}\|^2 + \alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^{x}\|^2 - \frac{1}{L_y}\|\boldsymbol{g}_{\star 1}^{y}\|^2.\end{aligned}$$

By triangle inequality and the Lipschitz gradient condition for L_{xy}, we have the following inequalities:

$$\begin{aligned}\|\boldsymbol{g}_{10}^{y}\|^2 - 2\|\boldsymbol{g}_{\star 0}^{y}\|^2 \le 2\|\boldsymbol{g}_{10}^{y} - \boldsymbol{g}_{\star 0}^{y}\|^2 \le 2L_{xy}^2\|\boldsymbol{x}_1\|^2,\\ \|\boldsymbol{g}_{11}^{x}\|^2 - 2\|\boldsymbol{g}_{1\star}^{x}\|^2 \le 2\|\boldsymbol{g}_{11}^{x} - \boldsymbol{g}_{1\star}^{x}\|^2 \le 2L_{xy}^2\|\boldsymbol{y}_1\|^2.\end{aligned}$$

If $\alpha \le \frac{1}{2L_x} \le \frac{1}{\sqrt{2}L_x}$ and $\beta \le \frac{1}{2L_y} \le \frac{1}{\sqrt{2}L_y}$, then we have

$$\begin{aligned}\alpha^2 L_x\|\boldsymbol{g}_{11}^{x}\|^2 - \frac{1}{L_x}\|\boldsymbol{g}_{1\star}^{x}\|^2 + \beta^2 L_y\|\boldsymbol{g}_{10}^{y}\|^2 - \frac{1}{L_y}\|\boldsymbol{g}_{\star 0}^{y}\|^2 &\le \alpha^2 L_x\left(\|\boldsymbol{g}_{11}^{x}\|^2 - 2\|\boldsymbol{g}_{1\star}^{x}\|^2\right) + \beta^2 L_y\left(\|\boldsymbol{g}_{10}^{y}\|^2 - 2\|\boldsymbol{g}_{\star 0}^{y}\|^2\right)\\ &\le 2\alpha^2 L_x L_{xy}^2\|\boldsymbol{y}_1\|^2 + 2\beta^2 L_y L_{xy}^2\|\boldsymbol{x}_1\|^2,\end{aligned}$$

and hence

$$\begin{aligned}\frac{1}{\alpha}\|\boldsymbol{x}_1\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2\|^2 \le{}& \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1\|^2\\ &+ 2\alpha^2 L_x L_{xy}^2\|\boldsymbol{y}_1\|^2 + 2\beta^2 L_y L_{xy}^2\|\boldsymbol{x}_1\|^2\\ &- \alpha(1-\alpha L_x)\|\boldsymbol{g}_{00}^{x}\|^2 + \alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^{x}\|^2 - \frac{1}{L_x}\|\boldsymbol{g}_{1\star}^{x}\|^2 - \frac{1}{L_y}\|\boldsymbol{g}_{\star 1}^{y}\|^2\\ \le{}& \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1\|^2\\ &+ 2\alpha^2 L_x L_{xy}^2\|\boldsymbol{y}_1\|^2 + 2\beta^2 L_y L_{xy}^2\|\boldsymbol{x}_1\|^2 - \alpha(1-\alpha L_x)\|\boldsymbol{g}_{00}^{x}\|^2 + \alpha(1-\alpha L_x)\|\boldsymbol{g}_{11}^{x}\|^2.\end{aligned}$$

Rearranging terms, we immediately have (39). □

D. Proofs used in Section 5

Here we prove all theorems related to **Alex-GDA** on SCSC Lipschitz gradient problems presented in Section 5.

- In Appendix D.1 we prove Theorem 5.1 which yields a contraction inequality for **Alex-GDA**.
- In Appendix D.2 we prove Corollary 5.2 which derives the corresponding iteration complexity upper bound.
- In Appendix D.3 we prove Theorem 5.3 which yields a matching lower bound for **Alex-GDA**.
- In Appendix D.4 we prove Proposition 5.4 which shows that the same lower bound holds for EG.
- In Appendix D.5 we prove technical propositions and lemmas used throughout the proofs in Appendix D.

D.1. Proof of Theorem 5.1

Here we prove Theorem 5.1 of Section 5, restated below for the sake of readability.

Theorem 5.1. *Suppose that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and we run **Alex-GDA** with $\gamma, \delta > 1$ and step sizes $\alpha, \beta > 0$ that satisfy*

$$\alpha \leq C \cdot \min\left\{\frac{1}{L_x}, \frac{\sqrt{\mu_y}}{L_{xy}\sqrt{\mu_x}}\right\},$$
$$\beta \leq C \cdot \min\left\{\frac{1}{L_y}, \frac{\sqrt{\mu_x}}{L_{xy}\sqrt{\mu_y}}\right\}.$$

for some constant $C > 0$ (which only depends on γ and δ). Then Ψ_k^{Alex} is valid, and satisfies $\Psi_{k+1}^{\text{Alex}} \leq r\Psi_k^{\text{Alex}}$ with

$$r = \max\left\{1 - \alpha\mu_x, 1 - \beta\mu_y\right\}.$$

Proof. Before starting the main proof, we characterize the step size condition as follows.

Finer Step Size Condition. We assume that the step sizes $\alpha, \beta > 0$ satisfy

$$\alpha \leq \frac{C_1}{L_x}, \quad \beta \leq \frac{C_2}{L_y}, \quad \alpha \leq \frac{C_3}{L_{xy}}\sqrt{\frac{\mu_y}{\mu_x}}, \quad \beta \leq \frac{C_4}{L_{xy}}\sqrt{\frac{\mu_x}{\mu_y}} \tag{47}$$

for constants $C_1, C_2, C_3, C_4 > 0$ satisfying

$$\begin{aligned} &C_1 \leq \frac{\gamma - 1}{2\gamma^2}, \quad C_2 \leq \frac{\delta - 1}{2\delta^2}, \\ &C_3 \leq \min\left\{\frac{1}{3\gamma - 2}, \frac{\delta - 1}{2(\gamma - 1)\delta}, \frac{1}{2(\gamma - 1)(\delta - 1)}\right\}, \\ &C_4 \leq \min\left\{\frac{1}{3\delta - 2}, \frac{\gamma - 1}{2\gamma(\delta - 1)}, \frac{1}{2(\gamma - 1)(\delta - 1)}\right\}. \end{aligned} \tag{48}$$

(By choosing $C = \min\{C_1, C_2, C_3, C_4\}$, we can obtain the simpler form given in the theorem statement.)

We show a few inequalities involving $C_1, C_2, C_3, C_4 > 0$ for future purposes.[9]

First, we have

$$C_1 \leq \frac{\gamma - 1}{2\gamma^2} \leq \frac{1}{2\gamma} \leq \frac{1}{2}, \quad C_2 \leq \frac{\delta - 1}{2\delta^2} \leq \frac{1}{2\delta} \leq \frac{1}{2}. \tag{49}$$

[9] Note that all arguments in the upper bounds of the constants given in (48) are all strictly positive whenever $\gamma > 1$ and $\delta > 1$.

Since $C_1 \leq \frac{\gamma-1}{2\gamma^2} \leq \frac{\gamma-1}{\gamma^2}$ and $C_4 \leq \frac{\gamma-1}{2\gamma(\delta-1)} \leq \frac{\gamma-1}{\gamma(\delta-1)}$, we have

$$\gamma^2 C_1 + \gamma(\delta-1)C_4 \leq 2(\gamma-1). \tag{50}$$

Since $C_1 \leq \frac{\gamma-1}{2\gamma^2} \leq \frac{1}{\gamma+1}$ and $C_4 \leq \frac{1}{3\delta-2}$, we have

$$(\gamma+1)C_1 + (3\delta-2)C_4 \leq 2. \tag{51}$$

Therefore, by (50) $+ (\gamma-1) \times$ (51) we have

$$(2\gamma^2-1)C_1 + (4\gamma\delta - 3\gamma - 3\delta + 2)\, C_4 \leq 4(\gamma-1). \tag{52}$$

Since $C_2 \leq \frac{\delta-1}{2\delta^2}$ and $C_3 \leq \frac{\delta-1}{2(\gamma-1)\delta}$, we have

$$\delta^2 C_2 + (\gamma-1)\delta C_3 \leq \delta - 1. \tag{53}$$

Since $C_2 \leq \frac{\delta-1}{2\delta^2} \leq \frac{1}{\delta+1}$ and $C_3 \leq \frac{1}{3\gamma-2}$, we have

$$(\delta+1)C_2 + (3\gamma-2)C_3 \leq 2. \tag{54}$$

Since $C_1 \leq \frac{1}{2}$ and $C_3 \leq \frac{1}{2(\gamma-1)(\delta-1)}$, we have

$$C_1 + (\gamma-1)(\delta-1)C_3 \leq 1, \tag{55}$$

and as $C_4 \leq \frac{1}{2(\gamma-1)(\delta-1)}$, we similarly have

$$C_1 + (\gamma-1)(\delta-1)C_4 \leq 1. \tag{56}$$

We also note that since $C_3 \leq \frac{\delta-1}{2(\gamma-1)\delta}$ and $C_4 \leq \frac{\gamma-1}{2\gamma(\delta-1)}$, we have

$$C_3 C_4 \leq \frac{1}{4\delta\gamma} \tag{57}$$

which, along with $\gamma, \delta > 1$, directly implies the followings:

$$4C_3C_4 \leq 1, \tag{58}$$
$$4(\delta-1)C_3C_4 \leq 1. \tag{59}$$

Now we proceed to the main proof of Theorem 5.1.

For $k \geq 1$, the Lyapunov function Ψ_k^{Alex} can be written as

$$\begin{aligned}\Psi_k^{\text{Alex}} = &\frac{1}{\alpha}\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_k - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)\|^2 \\ &+ (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2 + \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1}, \tilde{\boldsymbol{y}}_{k-1})\|^2,\end{aligned}$$

and for $k = 0$ as

$$\begin{aligned}\Psi_0^{\text{Alex}} = &\frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\ &+ \frac{(\gamma-1)(\delta-1)\alpha\beta}{(1-\alpha\mu_x)(1-\beta\mu_y)} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2.\end{aligned}$$

Similarly as in the proof of Theorem 4.1, the proof consists of two steps— in STEP 1 we prove that Ψ_k^{Alt} is a valid Lyapunov function, and in STEP 2 we show that $\Psi_{k+1}^{\text{Alt}} \leq r\Psi_k^{\text{Alt}}$ holds for the contraction rate r given as in Theorem 5.1.

STEP 1. VALIDITY OF LYAPUNOV FUNCTION

Here we show that there exists some constant A^{Alex} such that we have $\Psi_k^{\mathrm{Alex}} \geq A^{\mathrm{Alex}} \left(\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 + \|\boldsymbol{y}_k - \boldsymbol{y}_\star\|^2 \right)$, *i.e.*, Ψ_k^{Alex} is a valid Lyapunov function. Proposition D.1 yields a lower bound inequality from which we can derive such a constant A^{Alex}.

Proposition D.1. *Suppose that we run **Alex-GDA** with $\gamma, \delta > 0$ and step sizes α, β satisfying (47), and (48). Then we have*

$$\Psi_k^{\mathrm{Alex}} \geq \frac{1}{2\alpha}\|\boldsymbol{x}_k\|^2 + \frac{1}{2\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 \tag{60}$$

for all $(\boldsymbol{x}_k, \boldsymbol{y}_k)$, both when $k \geq 1$ and $k = 0$.

While we defer the proof of Proposition D.1 to Appendix D.5.1, here we see that this implies

$$\frac{1}{2\alpha}\|\boldsymbol{x}_k\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 \geq A^{\mathrm{Alex}} \left(\|\boldsymbol{x}_k\|^2 + \|\boldsymbol{y}_k\|^2 \right)$$

for $A^{\mathrm{Alex}} = \min\left\{ \frac{1}{2\alpha}, \frac{1}{\beta} \right\} > 0$.

STEP 2. CONTRACTION INEQUALITY

Note that this time we can't simply take $k = 0$ as in the proof of Theorem 4.1, since for **Alex-GDA** there exists a slight difference between the first iterate and the rest, as we have briefly explained in Section 5.

To deal with this subtlety, here we allow ourselves to set $k = 0$ W.L.O.G. by focusing on a set of iterates given by

$$\begin{aligned} \tilde{\boldsymbol{x}}_0 &= \boldsymbol{x}_0 - \xi(\gamma-1)\alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1}), \\ \tilde{\boldsymbol{y}}_0 &= \boldsymbol{y}_0 + \xi(\delta-1)\beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}), \end{aligned} \qquad \begin{aligned} \tilde{\boldsymbol{x}}_1 &= \boldsymbol{x}_0 - \gamma\alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0), \\ \boldsymbol{x}_1 &= \boldsymbol{x}_0 - \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0), \\ \tilde{\boldsymbol{y}}_1 &= \boldsymbol{y}_0 + \delta\beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0), \\ \boldsymbol{y}_1 &= \boldsymbol{y}_0 + \beta\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0), \\ \tilde{\boldsymbol{x}}_2 &= \boldsymbol{x}_1 - \gamma\alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1), \\ \boldsymbol{x}_2 &= \boldsymbol{x}_1 - \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1), \end{aligned} \tag{61}$$

where we can have either $\xi = 0$ or 1.

If $\xi = 0$, then we simply have $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0$ and $\boldsymbol{y}_0 = \tilde{\boldsymbol{y}}_0$, just as in the case of $k = 0$ of **Alex-GDA**. If $\xi = 1$, then we can bring the iterates $\tilde{\boldsymbol{x}}_0$ and $\tilde{\boldsymbol{y}}_0$ from the previous step, which corresponds to the case of $k \geq 1$ of **Alex-GDA**. Therefore it is safe to set $k = 0$ W.L.O.G., and it suffices to show a contraction inequality that holds for any iterates given by (61) (including both cases of $\xi = 0$ and 1), which we can apply to all iterates of the algorithm including both $k \geq 1$ and $k = 0$.

Proposition D.2 gives us the main inequality which leads to the desired contraction inequality.

Proposition D.2. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and iterates given by (61) with $\gamma, \delta > 0$ and step sizes α, β satisfying (47), and (48), we have the contraction inequality*

$$\begin{aligned} &\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 \\ &- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\ \leq &\left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 \\ &- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2, \end{aligned} \tag{62}$$

where $\xi = 0$ or 1.

We prove Proposition D.2 in Appendix D.5.2. Note that we can simplify the step size conditions as given in the theorem statement by choosing C as the minimum of the upper bounds of the constants given in (48).

First, let us assume that $\xi = 1$. Note that by Proposition D.2 we have

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + (\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + (\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2.
\end{aligned}$$

We can add $\frac{1}{1-\alpha\mu_x} \cdot (\gamma-1)(\delta-1)\alpha^2\beta L_{xy}\sqrt{\mu_x\mu_y}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2$ to both sides so that we have

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 \\
&+ (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\left(1 - \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} L_{xy}\sqrt{\mu_x\mu_y}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + (\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2.
\end{aligned} \tag{63}$$

Now let us define r as

$$r = \max\left\{1 - \alpha\mu_x, 1 - \beta\mu_y\right\}.$$

Since $r \geq 1 - \alpha\mu_x$ and $r \geq 1 - \beta\mu_y$, we have

$$\begin{aligned}
\left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 &\leq r \cdot \frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2, \\
2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 &\leq r \cdot \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2, \\
\left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 &\leq r \cdot \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2.
\end{aligned} \tag{64}$$

Since $r \geq 1 - \alpha\mu_x$, we have

$$(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 \leq r\frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2.$$

Now we will show that the following holds for the negative gradient terms:

$$-\alpha\left(1 - \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} L_{xy}\sqrt{\mu_x\mu_y}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \leq -r\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2. \tag{65}$$

Observe that

$$\frac{1}{1-\alpha\mu_x} \leq \frac{1}{1-\alpha L_x} \leq \frac{1}{1-C_1}. \tag{66}$$

Recalling inequality (56), we have

$$C_1 + (\gamma - 1)(\delta - 1)C_4 \leq 1,$$

which, combined with (66), gives

$$\frac{1}{1 - \alpha\mu_x}(\gamma - 1)(\delta - 1)C_4 \leq \frac{1}{1 - C_1}(\gamma - 1)(\delta - 1)C_4 \leq 1.$$

The condition $\beta \leq \frac{C_4}{L_{xy}}\sqrt{\frac{\mu_x}{\mu_y}}$ then yields

$$\frac{1 - \alpha\mu_x}{(\gamma - 1)(\delta - 1)\alpha\beta L_{xy}\sqrt{\mu_x\mu_y}} \geq \frac{1 - \alpha\mu_x}{(\gamma - 1)(\delta - 1)C_4\alpha\mu_x} \geq \frac{1}{\alpha\mu_x}.$$

Similarly, recalling inequality (55), we have

$$C_1 + (\gamma - 1)(\delta - 1)C_3 \leq 1,$$

which, combined with (66), gives

$$\frac{1}{1 - \alpha\mu_x}(\gamma - 1)(\delta - 1)C_3 \leq \frac{1}{1 - C_1}(\gamma - 1)(\delta - 1)C_3 \leq 1.$$

The conditions $\alpha \leq \frac{C_3}{L_{xy}}\sqrt{\frac{\mu_y}{\mu_x}}$ then yields

$$\frac{1 - \alpha\mu_x}{(\gamma - 1)(\delta - 1)\alpha\beta L_{xy}\sqrt{\mu_x\mu_y}} \geq \frac{1 - \alpha\mu_x}{(\gamma - 1)(\delta - 1)C_3\beta\mu_y} \geq \frac{1}{\beta\mu_y}.$$

Therefore we have

$$\frac{1 - \alpha\mu_x}{(\gamma - 1)(\delta - 1)\alpha\beta L_{xy}\sqrt{\mu_x\mu_y}} \geq \frac{1}{1 - r} = \max\left\{\frac{1}{\alpha\mu_x}, \frac{1}{\beta\mu_y}\right\},$$

or equivalently

$$1 - \frac{(\gamma - 1)(\delta - 1)\alpha\beta}{1 - \alpha\mu_x}L_{xy}\sqrt{\mu_x\mu_y} \geq r,$$

from which (65) immediately follows. Finally, we can just add:

$$0 \leq r(\delta - 1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2. \tag{67}$$

Aggregating (63), (64), (65), and (67), we can obtain

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 + (\delta - 1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \frac{(\gamma - 1)(\delta - 1)\alpha\beta}{1 - \alpha\mu_x}L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq r\left(\frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2\right) \\
&- r\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + r(\delta - 1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 + r\frac{(\gamma - 1)(\delta - 1)\alpha\beta}{1 - \alpha\mu_x}L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2
\end{aligned}$$

which – as $\xi = 1$ corresponds to iterates of **Alex-GDA** for $k \geq 1$ – concludes that $\Psi_{k+1}^{\text{Alex}} \leq r\Psi_k^{\text{Alex}}$ for $k \geq 1$.

Now suppose that $\xi = 0$. Then by Proposition D.2 we have

$$\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2$$
$$\leq \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2.$$

Then, since $1 - \beta\mu_y \leq r \leq 1$ and (64) holds for this case as well, we have

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq r\left(\frac{1}{\alpha}\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2\right) \\
&- r\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \frac{r}{1-\beta\mu_y} \cdot \frac{(\gamma-1)(\delta-1)\alpha\beta}{1-\alpha\mu_x} \cdot L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2
\end{aligned}$$

which – as $\xi = 0$ corresponds to the iterate of **Alex-GDA** for $k = 0$ – concludes that $\Psi_1^{\text{Alex}} \leq r\Psi_0^{\text{Alex}}$. □

D.2. Proof of Corollary 5.2

Here we prove Corollary 5.2 of Section 5, restated below for the sake of readability.

Corollary 5.2. *For step sizes given by the maximum possible values in Theorem 5.1, **Alex-GDA** linearly converges with iteration complexity*

$$\mathcal{O}\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{\Psi_0^{\text{Alex}}}{A^{\text{Alex}}\epsilon}\right),$$

where $A^{\text{Alex}} = \min\left\{\frac{1}{2\alpha}, \frac{1}{\beta}\right\} > 0$.

Proof. In Theorem 5.1 we have shown that $\Psi_{k+1}^{\text{Alex}} \leq r\Psi_k^{\text{Alex}}$ for all $k \geq 0$ with $r = \max\{1 - \alpha\mu_x, 1 - \beta\mu_y\}$.

Since we choose $\alpha = \Theta(\min\left\{\frac{1}{L_x}, \frac{\sqrt{\mu_y}}{L_{xy}\sqrt{\mu_x}}\right\})$ and $\beta = \Theta(\min\left\{\frac{1}{L_y}, \frac{\sqrt{\mu_x}}{L_{xy}\sqrt{\mu_y}}\right\})$, we have

$$\begin{aligned}
\frac{1}{1-r} &= \max\left\{\frac{1}{\alpha\mu_x}, \frac{1}{\beta\mu_y}\right\} \\
&= \Theta\left(\max\left\{\frac{L_x}{\mu_x}, \frac{L_y}{\mu_y}, \frac{L_{xy}}{\sqrt{\mu_x\mu_y}}\right\}\right) = \Theta\left(\kappa_x + \kappa_y + \kappa_{xy}\right).
\end{aligned}$$

Therefore it is sufficient to run

$$K = \mathcal{O}\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{\Psi_0^{\text{Alex}}}{A^{\text{Alex}}\epsilon}\right)$$

iterations to ensure that $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \leq \epsilon$, where $A^{\text{Alex}} = \min\left\{\frac{1}{2\alpha}, \frac{1}{\beta}\right\}$. □

D.3. Proof of Theorem 5.3

Here we prove Theorem 5.3 of Section 5, restated below for the sake of readability.

Theorem 5.3. *There exists a 6-dimensional function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ with $d_x = d_y = 3$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of Alex-GDA with $\gamma, \delta > 1$ requires an iteration complexity of*

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \le \epsilon$.

Proof. We use the same worst-case function as in Theorem 3.3:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} \begin{bmatrix} x \\ s \\ t \\ y \\ u \\ v \end{bmatrix}^\top \begin{bmatrix} \mu_x & 0 & 0 & L_{xy} & 0 & 0 \\ 0 & \mu_x & 0 & 0 & 0 & 0 \\ 0 & 0 & L_x & 0 & 0 & 0 \\ L_{xy} & 0 & 0 & -\mu_y & 0 & 0 \\ 0 & 0 & 0 & 0 & -\mu_y & 0 \\ 0 & 0 & 0 & 0 & 0 & -L_y \end{bmatrix} \begin{bmatrix} x \\ s \\ t \\ y \\ u \\ v \end{bmatrix},$$

where $\boldsymbol{x} = (x, s, t)$ and $\boldsymbol{y} = (y, u, v)$. It can be easily checked that f is a quadratic function (*i.e.,* Hessian is constant) such that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and $\boldsymbol{x}_\star = \boldsymbol{y}_\star = \boldsymbol{0} \in \mathbb{R}^3$.

We first observe that if we let

$$\boldsymbol{A} = \begin{bmatrix} \mu_x & 0 & 0 \\ 0 & \mu_x & 0 \\ 0 & 0 & L_x \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} L_{xy} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} \mu_y & 0 & 0 \\ 0 & \mu_y & 0 \\ 0 & 0 & L_y \end{bmatrix},$$

then the k-th step of Alex-GDA satisfies

$$\begin{aligned} \begin{bmatrix} \boldsymbol{x}_{k+1} \\ \tilde{\boldsymbol{x}}_{k+1} \\ \boldsymbol{y}_{k+1} \\ \tilde{\boldsymbol{y}}_{k+1} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \beta\boldsymbol{B}^\top & \boldsymbol{I} - \beta\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \delta\beta\boldsymbol{B}^\top & \boldsymbol{I} - \delta\beta\boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} - \alpha\boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & -\alpha\boldsymbol{B} \\ \boldsymbol{I} - \gamma\alpha\boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & -\gamma\alpha\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \tilde{\boldsymbol{x}}_k \\ \boldsymbol{y}_k \\ \tilde{\boldsymbol{y}}_k \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I} - \alpha\boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & -\alpha\boldsymbol{B} \\ \boldsymbol{I} - \gamma\alpha\boldsymbol{A} & \boldsymbol{0} & \boldsymbol{0} & -\gamma\alpha\boldsymbol{B} \\ \beta\boldsymbol{B}^\top(\boldsymbol{I} - \alpha\boldsymbol{A}) & \boldsymbol{0} & \boldsymbol{I} - \beta\boldsymbol{C} & -\gamma\alpha\beta\boldsymbol{B}^\top\boldsymbol{B} \\ \delta\beta\boldsymbol{B}^\top(\boldsymbol{I} - \gamma\alpha\boldsymbol{A}) & \boldsymbol{0} & \boldsymbol{I} - \delta\beta\boldsymbol{C} & -\gamma\delta\alpha\beta\boldsymbol{B}^\top\boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \tilde{\boldsymbol{x}}_k \\ \boldsymbol{y}_k \\ \tilde{\boldsymbol{y}}_k \end{bmatrix}. \end{aligned}$$

Therefore we have the following coordinate-wise updates:

$$\begin{bmatrix} x_{k+1} \\ \tilde{x}_{k+1} \\ y_{k+1} \\ \tilde{y}_{k+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 - \alpha\mu_x & 0 & 0 & -\alpha L_{xy} \\ 1 - \gamma\alpha\mu_x & 0 & 0 & -\gamma\alpha L_{xy} \\ \beta L_{xy}(1 - \gamma\alpha\mu_x) & 0 & 1 - \beta\mu_y & -\gamma\alpha\beta L_{xy}^2 \\ \delta\beta L_{xy}(1 - \gamma\alpha\mu_x) & 0 & 1 - \delta\beta\mu_y & -\gamma\delta\alpha\beta L_{xy}^2 \end{bmatrix}}_{\triangleq \boldsymbol{P}} \begin{bmatrix} x_k \\ \tilde{x}_k \\ y_k \\ y_k \end{bmatrix}, \tag{68}$$

$$s_{k+1} = (1 - \alpha\mu_x)s_k, \quad \tilde{s}_{k+1} = (1 - \gamma\alpha\mu_x)s_k, \tag{69}$$

$$t_{k+1} = (1 - \alpha L_x)t_k, \quad \tilde{t}_{k+1} = (1 - \gamma\alpha L_x)t_k, \tag{70}$$

$$u_{k+1} = (1 - \beta\mu_y)u_k, \quad \tilde{u}_{k+1} = (1 - \delta\beta\mu_y)u_k, \tag{71}$$

$$v_{k+1} = (1 - \beta L_y)v_k, \quad \tilde{v}_{k+1} = (1 - \delta\beta L_y)v_k. \tag{72}$$

To assure the convergence of iterations (70) and (72), the step sizes α and β are required to be

$$\alpha < \frac{2}{L_x} \quad \text{and} \quad \beta < \frac{2}{L_y}. \tag{73}$$

Also, to guarantee $\|\boldsymbol{x}_K\|^2 + \|\boldsymbol{y}_K\|^2 < \epsilon$, we need from (69) and (71) that $s_K^2 < \mathcal{O}(\epsilon)$ and $u_K^2 < \mathcal{O}(\epsilon)$, respectively. These two necessary conditions require an iteration number of at least:

$$K = \Omega\left(\left(\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y}\right) \cdot \log\frac{1}{\epsilon}\right), \tag{74}$$

and $\alpha L_x, \beta L_y = \mathcal{O}(1)$ from (74) yields

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y). \tag{75}$$

Now, in order to ensure convergence of iteration (68), we need the following matrix

$$\boldsymbol{P} = \begin{bmatrix} 1-\alpha\mu_x & 0 & 0 & -\alpha L_{xy} \\ 1-\gamma\alpha\mu_x & 0 & 0 & -\gamma\alpha L_{xy} \\ \beta L_{xy}(1-\gamma\alpha\mu_x) & 0 & 1-\beta\mu_y & -\gamma\alpha\beta L_{xy}^2 \\ \delta\beta L_{xy}(1-\gamma\alpha\mu_x) & 0 & 1-\delta\beta\mu_y & -\gamma\delta\alpha\beta L_{xy}^2 \end{bmatrix}$$

to have a spectral radius smaller than one. Hence it suffices to show that $\rho(\boldsymbol{P}) < 1$ implies that $\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_{xy})$.

Suppose that λ is an eigenvalue of $\boldsymbol{P}$. Then we must have

$$\begin{aligned}\det(\lambda\boldsymbol{I} - \boldsymbol{P}) &= \begin{vmatrix} (1-\lambda)-\alpha\mu_x & 0 & 0 & -\alpha L_{xy} \\ 1-\gamma\alpha\mu_x & -\lambda & 0 & -\gamma\alpha L_{xy} \\ \beta L_{xy}(1-\gamma\alpha\mu_x) & 0 & (1-\lambda)-\beta\mu_y & -\gamma\alpha\beta L_{xy}^2 \\ \delta\beta L_{xy}(1-\gamma\alpha\mu_x) & 0 & 1-\delta\beta\mu_y & -\lambda-\gamma\delta\alpha\beta L_{xy}^2 \end{vmatrix} \\ &= -\lambda \cdot \begin{vmatrix} (1-\lambda)-\alpha\mu_x & 0 & -\alpha L_{xy} \\ \beta L_{xy}(1-\gamma\alpha\mu_x) & (1-\lambda)-\beta\mu_y & -\gamma\alpha\beta L_{xy}^2 \\ \delta\beta L_{xy}(1-\gamma\alpha\mu_x) & 1-\delta\beta\mu_y & -\lambda-\gamma\delta\alpha\beta L_{xy}^2 \end{vmatrix} = 0.\end{aligned}$$

We can compute

$$\begin{aligned}&\begin{vmatrix} (1-\lambda)-\alpha\mu_x & 0 & -\alpha L_{xy} \\ \beta L_{xy}(1-\gamma\alpha\mu_x) & (1-\lambda)-\beta\mu_y & -\gamma\alpha\beta L_{xy}^2 \\ \delta\beta L_{xy}(1-\gamma\alpha\mu_x) & 1-\delta\beta\mu_y & -\lambda-\gamma\delta\alpha\beta L_{xy}^2. \end{vmatrix} \\ &= ((1-\lambda)-\alpha\mu_x)\left((1-\lambda)-\beta\mu_y\right)\left(-\lambda-\gamma\delta\alpha\beta L_{xy}^2\right) - \alpha\beta L_{xy}^2(1-\gamma\alpha\mu_x)(1-\delta\beta\mu_y) \\ &\quad + \delta\alpha\beta L_{xy}^2(1-\gamma\alpha\mu_x)\left((1-\lambda)-\beta\mu_y\right) + \gamma\alpha\beta L_{xy}^2(1-\delta\beta\mu_y)\left((1-\lambda)-\alpha\mu_x\right).\end{aligned}$$

Substituting $\lambda = 1 - t$ and $\phi = \alpha\beta L_{xy}^2$, we can obtain a simpler expression:

$$\begin{aligned}&(t-\alpha\mu_x)(t-\beta\mu_y)(t-1-\gamma\delta\phi) - \phi(1-\gamma\alpha\mu_x)(1-\delta\beta\mu_y) \\ &\quad + \delta\phi(1-\gamma\alpha\mu_x)(t-\beta\mu_y) + \gamma\phi(1-\delta\beta\mu_y)(t-\alpha\mu_x) \\ &= (t-\alpha\mu_x)(t-\beta\mu_y)(t-1) - \gamma\delta\phi(t-\alpha\mu_x)(t-\beta\mu_y) - \phi(1-\gamma\alpha\mu_x)(1-\delta\beta\mu_y) \\ &\quad + \delta\phi(1-\gamma\alpha\mu_x)(t-\beta\mu_y) + \gamma\phi(1-\delta\beta\mu_y)(t-\alpha\mu_x) \\ &= (t-\alpha\mu_x)(t-\beta\mu_y)(t-1) - \phi\left((1-\gamma\alpha\mu_x) - \gamma(t-\alpha\mu_x)\right)\left((1-\delta\beta\mu_y) - \delta(t-\beta\mu_y)\right) \\ &= (t-\alpha\mu_x)(t-\beta\mu_y)(t-1) - \phi(1-\gamma t)(1-\delta t).\end{aligned}$$

Therefore the eigenvalue λ must be 0 or take the form of $1 - t^*$, where t^* is a root of the following cubic equation:

$$(t-\alpha\mu_x)(t-\beta\mu_y)(t-1) - \phi(1-\gamma t)(1-\delta t) = 0.$$

We can expand as

$$t^3 - (1 + \alpha\mu_x + \beta\mu_y + \gamma\delta\phi)t^2 + (\alpha\mu_x + \beta\mu_y + \alpha\beta\mu_x\mu_y + (\gamma+\delta)\phi)t - (\alpha\beta\mu_x\mu_y + \phi) = 0.$$

Hence we have a cubic equation of the form $t^3 - pt^2 + qt - r = 0$ with coefficients given by

$$\begin{aligned} p &= 1 + \alpha\mu_x + \beta\mu_y + \gamma\delta\phi, \\ q &= \alpha\mu_x + \beta\mu_y + \alpha\beta\mu_x\mu_y + (\gamma+\delta)\phi, \\ r &= \alpha\beta\mu_x\mu_y + \phi. \end{aligned} \tag{76}$$

Note that we obviously have $p, q, r > 0$.

There exists a well-known characterization of cubic polynomials having roots with absolute values less than one.

Proposition D.3 (Grove & Ladas (2004), Theorem 1.4)**.** *Consider a cubic polynomial $x^3 + a_2x^2 + a_1x + a_0$, where a_0, a_1, and a_2 are real numbers. Then a necessary and sufficient condition that all roots of the polynomial are contained in the open disk $|x| < 1$ is*

$$|a_2 + a_0| < 1 + a_1, \quad |a_2 - 3a_0| < 3 - a_1, \quad a_0(a_0 - a_2) + a_1 - 1 < 0. \tag{77}$$

Also, the following corollary suggests that the coefficients are all bounded (by constants) for such cases.

Corollary D.4. *For coefficients a_0, a_1, a_2 satisfying (77), we have $|a_2| < 3$, $|a_1| < 3$, and $|a_0| < 1$.*

Proof. It is easy to see that $-1 < a_1 < 3$ from the first two conditions.

Also, the first and the last condition together imply that

$$|a_2 + a_0| - 1 < a_1 < a_0(a_2 - a_0) + 1.$$

This is a subset of the region

$$|a_2 + a_0| < 4 \quad \wedge \quad |a_2 + a_0| < a_0(a_2 - a_0) + 2.$$

The range of such (a_2, a_0) is equal to a parallelogram with endpoints $(-3, -1)$, $(1, -1)$, $(-1, 1)$, $(3, 1)$, which implies $|a_2| < 3$ and $|a_0| < 1$. □

Plugging back in $t = 1 - \lambda$, we can write the cubic polynomial in terms of p, q, r, and λ as

$$\begin{aligned} & (1-\lambda)^3 - p(1-\lambda)^2 + q(1-\lambda) - r = 0 \\ \Leftrightarrow \quad & \lambda^3 + (-3+p)\lambda^2 + (3-2p+q)\lambda + (-1+p-q+r) = 0. \end{aligned}$$

By Corollary D.4, we can observe that a necessary condition for $\rho(\boldsymbol{P}) < 1$ is that

$$|3 - p| < 3, \;\; |3 - 2p + q| < 3, \;\; |1 - p + q - r| < 1.$$

We can simply deduce that $p < 6$, which implies $q < 12$ and finally $r < 14$.

Therefore we can conclude that all of the coefficients in (76) are of order $\mathcal{O}(1)$. In particular, this implies $\phi = \alpha\beta L_{xy}^2 = \mathcal{O}(1)$ in order to assure convergence, which concludes that

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} \geq \frac{2}{\sqrt{\alpha\beta\mu_x\mu_y}} = \frac{2\kappa_{xy}}{\sqrt{\alpha\beta L_{xy}^2}} = \Omega(\kappa_{xy}). \tag{78}$$

Combining (75) and (78), we have

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y + \kappa_{xy})$$

and therefore from (74) we can show a lower bound of

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log\frac{1}{\epsilon}\right).$$

□

D.4. Proof of Proposition 5.4

Here we prove Proposition 5.4 of Section 5, restated below for the sake of readability.

Proposition 5.4. *There exists a* 6*-dimensional function* $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ *with* $d_x = d_y = 3$ *such that for any constant step sizes* $\alpha, \beta > 0$*, the convergence of* ***EG*** *requires an iteration complexity of rate at least*

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \le \epsilon$.

Proof. Recall that **EG** takes updates of the form:

$$\begin{aligned}
\boldsymbol{x}_{k+\frac{1}{2}} &= \boldsymbol{x}_k - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \\
\boldsymbol{y}_{k+\frac{1}{2}} &= \boldsymbol{y}_k + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \\
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}), \\
\boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}).
\end{aligned}$$

We use the same worst-case function as in Theorem 3.3:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} \begin{bmatrix} x \\ s \\ t \\ y \\ u \\ v \end{bmatrix}^\top \begin{bmatrix} \mu_x & 0 & 0 & L_{xy} & 0 & 0 \\ 0 & \mu_x & 0 & 0 & 0 & 0 \\ 0 & 0 & L_x & 0 & 0 & 0 \\ L_{xy} & 0 & 0 & -\mu_y & 0 & 0 \\ 0 & 0 & 0 & 0 & -\mu_y & 0 \\ 0 & 0 & 0 & 0 & 0 & -L_y \end{bmatrix} \begin{bmatrix} x \\ s \\ t \\ y \\ u \\ v \end{bmatrix},$$

where $\boldsymbol{x} = (x, s, t)$ and $\boldsymbol{y} = (y, u, v)$. It can be easily checked that f is a quadratic function (*i.e.,* Hessian is constant) such that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and $\boldsymbol{x}_\star = \boldsymbol{y}_\star = \boldsymbol{0} \in \mathbb{R}^3$.

Let us define

$$\boldsymbol{A} = \begin{bmatrix} \mu_x & 0 & 0 \\ 0 & \mu_x & 0 \\ 0 & 0 & L_x \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} L_{xy} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} \mu_y & 0 & 0 \\ 0 & \mu_y & 0 \\ 0 & 0 & L_y \end{bmatrix}.$$

We first observe that the k-th step of **EG** satisfies

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{x}_{k+\frac{1}{2}} \\ \boldsymbol{y}_{k+\frac{1}{2}} \end{bmatrix} &= \underbrace{\begin{bmatrix} \boldsymbol{I} - \alpha \boldsymbol{A} & -\alpha \boldsymbol{B} \\ \beta \boldsymbol{B}^\top & \boldsymbol{I} - \beta \boldsymbol{C} \end{bmatrix}}_{\triangleq \boldsymbol{M}_{\text{Sim}}} \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix}, \\
\begin{bmatrix} \boldsymbol{x}_{k+1} \\ \boldsymbol{y}_{k+1} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix} + \begin{bmatrix} -\alpha \boldsymbol{A} & -\alpha \boldsymbol{B} \\ \beta \boldsymbol{B}^\top & -\beta \boldsymbol{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{k+\frac{1}{2}} \\ \boldsymbol{y}_{k+\frac{1}{2}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix} + \begin{bmatrix} -\alpha \boldsymbol{A} & -\alpha \boldsymbol{B} \\ \beta \boldsymbol{B}^\top & -\beta \boldsymbol{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} - \alpha \boldsymbol{A} & -\alpha \boldsymbol{B} \\ \beta \boldsymbol{B}^\top & \boldsymbol{I} - \beta \boldsymbol{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix} \\
&= (\boldsymbol{I} + (\boldsymbol{M}_{\text{Sim}} - \boldsymbol{I}) \boldsymbol{M}_{\text{Sim}}) \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix} = \underbrace{(\boldsymbol{I} - \boldsymbol{M}_{\text{Sim}} + \boldsymbol{M}_{\text{Sim}}^2)}_{\triangleq \boldsymbol{M}_{\text{EG}}} \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix}.
\end{aligned}$$

Hence we have that λ_{Sim} is an eigenvalue of $\boldsymbol{M}_{\text{Sim}}$ if and only if $\lambda_{\text{EG}} = 1 - \lambda_{\text{Sim}} + \lambda_{\text{Sim}}^2$ is an eigenvalue of $\boldsymbol{M}_{\text{EG}}$. Note that the matrix $\boldsymbol{M}_{\text{Sim}}$ is identical to the updates made by **Sim-GDA** on the same lower bound function f, which allows us to utilize some results from Appendix B.3.

Let us define

$$\boldsymbol{P} \triangleq \begin{bmatrix} 1-\alpha\mu_x & -\alpha L_{xy} \\ \beta L_{xy} & 1-\beta\mu_y \end{bmatrix}.$$

Then the k-th step of **EG** satisfies

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{P}^2) \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \tag{79}$$

$$s_{k+1} = (1 - \alpha\mu_x + \alpha^2\mu_x^2)s_k, \tag{80}$$

$$t_{k+1} = (1 - \alpha L_x + \alpha^2 L_x^2)t_k, \tag{81}$$

$$u_{k+1} = (1 - \beta\mu_y + \beta^2\mu_y^2)u_k, \tag{82}$$

$$v_{k+1} = (1 - \beta L_y + \beta^2 L_y^2)v_k. \tag{83}$$

We can see that the eigenvalues of $\boldsymbol{M}_{\text{EG}}$ must be either $\lambda_{\text{EG}} = 1 - \lambda_P + \lambda_P^2$, where λ_P is an eigenvalue of $\boldsymbol{P}$, which can be explicitly computed as

$$\lambda_P = 1 - \frac{\alpha\mu_x + \beta\mu_y}{2} \pm \sqrt{\left(\frac{\alpha\mu_x - \beta\mu_y}{2}\right)^2 - \alpha\beta L_{xy}^2} \tag{84}$$

or among the following values:

$$1 - \alpha\mu_x, \;\; 1 - \alpha L_x, \;\; 1 - \beta\mu_y, \;\; \text{and} \;\; 1 - \beta L_y. \tag{85}$$

For the (real) eigenvalues in (85), we can deduce that the corresponding eigenvalues of $\boldsymbol{M}_{\text{EG}}$ are

$$1 - \alpha\mu_x + \alpha^2\mu_x^2, \;\; 1 - \alpha L_x + \alpha^2 L_x^2, \;\; 1 - \beta\mu_y + \beta^2\mu_y^2, \;\; \text{and} \;\; 1 - \beta L_y + \beta^2 L_y^2,$$

all being strictly larger than the corresponding values in (85). Hence, for the convergence of iterations (81) and (83), the step sizes α and β are required to satisfy

$$0 < \alpha L_x(1 - \alpha L_x) < 2 \quad \text{and} \quad 0 < \beta L_y(1 - \beta L_y) < 2,$$

which (as $\alpha, \beta > 0$) is simply equivalent to

$$\alpha < \frac{1}{L_x} \quad \text{and} \quad \beta < \frac{1}{L_y}. \tag{86}$$

Also, to guarantee $\|\boldsymbol{x}_K\|^2 + \|\boldsymbol{y}_K\|^2 < \epsilon$, we need from (80) and (82) that $s_K^2 < \mathcal{O}(\epsilon)$ and $u_K^2 < \mathcal{O}(\epsilon)$, respectively. These two necessary conditions require an iteration number of at least:

$$K = \Omega\left(\left(\frac{1}{\alpha\mu_x(1-\alpha\mu_x)} + \frac{1}{\beta\mu_y(1-\beta\mu_y)}\right) \cdot \log\frac{1}{\epsilon}\right) = \Omega\left(\left(\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y}\right) \cdot \log\frac{1}{\epsilon}\right). \tag{87}$$

Note that (86) automatically yields

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y). \tag{88}$$

Now we focus on the x, y coordinates to complete the proof. We do a similar case-by-case analysis as in our proof of Theorem 3.3 in Appendix B.3, based on whether the eigenvalues in (84) are real or complex.

Case 1. If the eigenvalues λ_P in (84) are real, then we have

$$\left|\sqrt{\frac{\alpha\mu_x}{\beta\mu_y}} - \sqrt{\frac{\beta\mu_y}{\alpha\mu_x}}\right| > 2\kappa_{xy}$$

as in (23) of Appendix B.3. By the same logic as in **Case 2** of Appendix B.3, we have

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y + \kappa_{xy}^2) = \Omega(\kappa_x + \kappa_y + \kappa_{xy}). \tag{89}$$

Case 2. Suppose that the eigenvalues in (84) are complex. If we substitute as

$$s = \frac{\alpha\mu_x + \beta\mu_y}{2}, \;\; p = \sqrt{\alpha\beta\mu_x\mu_y}, \;\; K = \kappa_{xy}^2,$$

then (84) can be written as:

$$\lambda_P = 1 - s \pm i\sqrt{(K+1)p^2 - s^2}.$$

As we consider the case when the eigenvalues are complex, here we must have

$$s^2 \leq (K+1)p^2.$$

We can explicitly compute

$$\begin{aligned}
\lambda_{\text{EG}} &= \lambda_P + (1-\lambda_P)^2 \\
&= 1 - s \pm i\sqrt{(K+1)p^2 - s^2} + \left(s \mp i\sqrt{(K+1)p^2 - s^2}\right)^2 \\
&= 1 - s + s^2 - ((K+1)p^2 - s^2) \pm i\left((1-2s)\sqrt{(K+1)p^2 - s^2}\right)
\end{aligned}$$

Therefore $|\lambda_{\text{EG}}|^2$ can be expressed as

$$\begin{aligned}
&\left(1 - s + s^2 - ((K+1)p^2 - s^2)\right)^2 + (1-2s)^2((K+1)p^2 - s^2) \\
&= (1 - s + s^2)^2 - 2(1 - s + s^2)((K+1)p^2 - s^2) + ((K+1)p^2 - s^2)^2 + (1-2s)^2((K+1)p^2 - s^2) \\
&= (1 - s + s^2)^2 - (2(1 - s + s^2) - (1-2s)^2)((K+1)p^2 - s^2) + ((K+1)p^2 - s^2)^2 \\
&= (1 - s + s^2)^2 - (1 + 2s - 2s^2)((K+1)p^2 - s^2) + ((K+1)p^2 - s^2)^2 \\
&= (1 - s + s^2)^2 + ((K+1)p^2 + s^2 - 2s - 1)((K+1)p^2 - s^2) \\
&= (1 - s + s^2)^2 + (K+1)^2p^4 - s^4 - (2s+1)((K+1)p^2 - s^2) \\
&= (1 - s + s^2)^2 + (2s+1)s^2 - s^4 - (K+1)p^2(2s+1) + (K+1)^2p^4 \\
&= 1 - 2s + 4s^2 - (K+1)p^2(2s+1) + (K+1)^2p^4.
\end{aligned}$$

Note that $|\lambda_{\text{EG}}| < 1$ is equivalent to

$$-(K+1)p^2(2s+1) + (K+1)^2p^4 < 2s - 4s^2.$$

If this is true, then substituting $t = (K+1)p^2$ we obtain the following region:

$$t^2 - t(2s+1) + 4s^2 - 2s < 0, \;\; s^2 \leq t$$

which is the upper region of the interior of an ellipse cut by a parabola. This region is bounded, and we can compute the range of t as $0 < t < 1 + 2/\sqrt{3}$. Therefore we have $t = \mathcal{O}(1)$, and since we can substitute back as

$$t = (K+1)p^2 = (\kappa_{xy}^2 + 1)\alpha\beta\mu_x\mu_y,$$

we can observe that $(\kappa_{xy}^2 + 1)\alpha\beta\mu_x\mu_y = \mathcal{O}(1)$, and therefore

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} \geq \frac{2}{\sqrt{\alpha\beta\mu_x\mu_y}} = \frac{2\sqrt{(\kappa_{xy}^2+1)}}{\sqrt{(\kappa_{xy}^2+1)\alpha\beta\mu_x\mu_y}} = \Omega(\kappa_{xy}).$$

Aggregating with (88), we can observe that

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y + \kappa_{xy}),$$

and hence the lower bound iteration complexity holds for all possible cases of convergence. □

D.5. Proofs used in Appendix D

Here we prove some technical propositions and lemmas used throughout Appendix D.

D.5.1. Proof of Proposition D.1

Here we prove Proposition D.1, restated below for the sake of readability.

Proposition D.1. *Suppose that we run Alex-GDA with $\gamma, \delta > 0$ and step sizes α, β satisfying (47), and (48). Then we have*

$$\Psi_k^{\mathrm{Alex}} \geq \frac{1}{2\alpha}\|\boldsymbol{x}_k\|^2 + \frac{1}{2\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 \tag{60}$$

for all $(\boldsymbol{x}_k, \boldsymbol{y}_k)$, both when $k \geq 1$ and $k = 0$.

Proof. For simplicity let us assume W.L.O.G. that $\boldsymbol{x}_\star = \mathbf{0}$ $(\in \mathbb{R}^{d_x})$ and $\boldsymbol{y}_\star = \mathbf{0}$ $(\in \mathbb{R}^{d_y})$.

For $k \geq 1$, we have

$$\Psi_k^{\mathrm{Alex}} \geq \frac{1}{\alpha}\|\boldsymbol{x}_k\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2.$$

By triangle inequality and Lipschitz gradients, we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)\|^2 &\leq 2\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_\star, \tilde{\boldsymbol{y}}_k)\|^2 + 2\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_\star, \tilde{\boldsymbol{y}}_k)\|^2 \\
&\leq 2L_x^2\|\boldsymbol{x}_k\|^2 + 2L_{xy}^2\|\tilde{\boldsymbol{y}}_k\|^2 \\
&\leq 2L_x^2\|\boldsymbol{x}_k\|^2 + 4L_{xy}^2\|\boldsymbol{y}_k\|^2 + 4L_{xy}^2\|\tilde{\boldsymbol{y}}_k - \boldsymbol{y}_k\|^2 \\
&\leq 2L_x^2\|\boldsymbol{x}_k\|^2 + 4L_{xy}^2\|\boldsymbol{y}_k\|^2 + 4(\delta-1)^2\beta^2 L_{xy}^2\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2
\end{aligned}$$

Therefore, we can obtain

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_k\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2 \\
&\geq \left(\frac{1}{\alpha} - 2\alpha L_x^2\right)\|\boldsymbol{x}_k\|^2 + 2\left(\frac{1}{\beta} - 2\alpha L_{xy}^2\right)\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 \\
&\quad + (\delta-1)\beta\left(1 - 4(\delta-1)\alpha\beta L_{xy}^2\right)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2.
\end{aligned}$$

Since $\alpha \leq \frac{C_1}{L_x}$ and $C_1 \leq \frac{1}{2}$ (by (49)), we have

$$\frac{1}{\alpha} - 2\alpha L_x^2 \geq \frac{1}{\alpha} - \frac{2C_1^2}{\alpha} \geq \frac{1}{2\alpha}.$$

Since $\alpha \leq \frac{C_3}{L_{xy}}\sqrt{\frac{\mu_y}{\mu_x}}$, $\beta \leq \frac{C_4}{L_{xy}}\sqrt{\frac{\mu_x}{\mu_y}}$, and $4C_3C_4 \leq 1$ (by (58)), we have

$$\frac{1}{\beta} - 2\alpha L_{xy}^2 \geq \frac{1}{\beta} - \frac{2C_3C_4}{\beta}\sqrt{\frac{\mu_y}{\mu_x}}\cdot\sqrt{\frac{\mu_x}{\mu_y}} = \frac{1}{\beta} - \frac{2C_3C_4}{\beta} \geq \frac{1}{2\beta}.$$

Finally, since $4(\delta-1)C_3C_4 \leq 1$ (by (59)), we have

$$4(\delta-1)\alpha\beta L_{xy}^2 \leq 4(\delta-1)C_3C_4 \leq 1$$

and therefore we can cancel out the last term by

$$(\delta-1)\beta\left(1 - 4(\delta-1)\alpha\beta L_{xy}^2\right)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2 \geq 0.$$

Therefore we have

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_k\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_k, \boldsymbol{y}_{k-1})\|^2 \\
&\geq \frac{1}{2\alpha}\|\boldsymbol{x}_k\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_k\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_{k+1}\|^2,
\end{aligned}$$

which implies that (60) is indeed true for $k = 1$.

For $k = 0$, we have

$$\Psi_0^{\text{Alex}} \geq \frac{1}{\alpha}\|\boldsymbol{x}_0\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0)\|^2,$$

where we note that $\tilde{\boldsymbol{y}}_0 = \boldsymbol{y}_0$. By triangle inequality and Lipschitz gradients, we have

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0)\|^2 \leq 2\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_\star, \boldsymbol{y}_0)\|^2 + 2\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_\star, \boldsymbol{y}_0)\|^2 \leq 2L_x^2\|\boldsymbol{x}_0\|^2 + 2L_{xy}^2\|\boldsymbol{y}_0\|^2.$$

Therefore, we can obtain

$$\frac{1}{\alpha}\|\boldsymbol{x}_0\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0)\|^2 \geq \left(\frac{1}{\alpha} - 2\alpha L_x^2\right)\|\boldsymbol{x}_0\|^2 + 2\left(\frac{1}{\beta} - \alpha L_{xy}^2\right)\|\boldsymbol{y}_0\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1\|^2.$$

Since $\alpha \leq \frac{C_1}{L_x}$ and $C_1 \leq \frac{1}{2}$ (by (49)), we have

$$\frac{1}{\alpha} - 2\alpha L_x^2 \geq \frac{1}{\alpha} - \frac{2C_1^2}{\alpha} \geq \frac{1}{2\alpha}.$$

Since $\alpha \leq \frac{C_3}{L_{xy}}\sqrt{\frac{\mu_y}{\mu_x}}$, $\beta \leq \frac{C_4}{L_{xy}}\sqrt{\frac{\mu_x}{\mu_y}}$, and $C_3C_4 \leq \frac{1}{4}$ (by (58)), we have

$$\frac{1}{\beta} - 2\alpha L_{xy}^2 \geq \frac{1}{\beta} - \frac{2C_3C_4}{\beta}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \sqrt{\frac{\mu_x}{\mu_y}} = \frac{1}{\beta} - \frac{2C_3C_4}{\beta} \geq \frac{1}{2\beta}.$$

Therefore we have

$$\frac{1}{\alpha}\|\boldsymbol{x}_0\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_0\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0)\|^2 \geq \frac{1}{2\alpha}\|\boldsymbol{x}_0\|^2 + \frac{1}{\beta}\|\boldsymbol{y}_0\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_1\|^2,$$

which implies that (60) is indeed true for $k = 0$. □

D.5.2. Proof of Proposition D.2

Here we prove Proposition D.2, restated below for the sake of readability.

Proposition D.2. *For $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and iterates given by (61) with $\gamma, \delta > 0$ and step sizes α, β satisfying (47), and (48), we have the contraction inequality*

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 + (\delta - 1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \xi(\gamma - 1)(\delta - 1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 \\
&- \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma - 1)(\delta - 1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2,
\end{aligned} \tag{62}$$

where $\xi = 0$ or 1.

Proof. While the proof of the proposition is quite technical and complicated, we can largely divide the proof into three large steps. In Step 1, we use the basic notions of strong convexity (and/or strong concavity) and the Lipschitz gradient conditions involving L_x and L_y (*i.e., smoothness* in convex optimization literature) to obtain an inequality between terms from the previous and next iterates. In Step 2, we use the L_{xy}-Lipschitz gradient conditions to cope with the intermediate inner product terms. In Step 3, we use the given step size conditions to cancel out the gradient norm terms as much as possible, which leaves us with the inequality given in the proposition statement.

STEP 1. BASIC TRANSFORMATIONS

We start with

$$\begin{aligned}
\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 &= \frac{1}{\alpha}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+\frac{2}{\alpha}\left\langle \boldsymbol{x}_1-\boldsymbol{x}_0,\boldsymbol{x}_0-\boldsymbol{x}_\star\right\rangle+\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2\\
&= \frac{1}{\alpha}\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2-2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0),\boldsymbol{x}_0-\boldsymbol{x}_\star\right\rangle+\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2,\\
\frac{2}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2 &= \frac{2}{\beta}\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+\frac{4}{\beta}\left\langle \boldsymbol{y}_1-\boldsymbol{y}_0,\boldsymbol{y}_0-\boldsymbol{y}_\star\right\rangle+\frac{2}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_0\|^2\\
&= \frac{2}{\beta}\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+4\left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0),\boldsymbol{y}_0-\boldsymbol{y}_\star\right\rangle+2\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2,\\
\frac{1}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2 &= \frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+\frac{2}{\alpha}\left\langle \boldsymbol{x}_2-\boldsymbol{x}_1,\boldsymbol{x}_1-\boldsymbol{x}_\star\right\rangle+\frac{1}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_1\|^2\\
&= \frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2-2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1),\boldsymbol{x}_1-\boldsymbol{x}_\star\right\rangle+\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)\|^2.
\end{aligned}\tag{90}$$

By strong convexity (concavity), we have

$$\begin{aligned}
-2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0),\boldsymbol{x}_0-\boldsymbol{x}_\star\right\rangle &\le -\mu_x\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2-2(f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)-f(\boldsymbol{x}_\star,\tilde{\boldsymbol{y}}_0)),\\
4\left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0),\boldsymbol{y}_0-\boldsymbol{y}_\star\right\rangle &\le -2\mu_y\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+4(f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)-f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_\star)),\\
-2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1),\boldsymbol{x}_1-\boldsymbol{x}_\star\right\rangle &\le -\mu_x\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2-2(f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)-f(\boldsymbol{x}_\star,\tilde{\boldsymbol{y}}_1)).
\end{aligned}\tag{91}$$

Since f has Lipschitz gradients, we have

$$\begin{aligned}
2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0),\boldsymbol{x}_0-\tilde{\boldsymbol{x}}_1\right\rangle &\le L_x\|\boldsymbol{x}_0-\tilde{\boldsymbol{x}}_1\|^2+2(f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_0)),\\
-2\left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0),\boldsymbol{y}_0-\tilde{\boldsymbol{y}}_0\right\rangle &\le L_y\|\boldsymbol{y}_0-\tilde{\boldsymbol{y}}_0\|^2-2(f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_0)),\\
-2\left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0),\boldsymbol{y}_0-\tilde{\boldsymbol{y}}_1\right\rangle &\le L_y\|\boldsymbol{y}_0-\tilde{\boldsymbol{y}}_1\|^2-2(f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_1)),\\
2\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1),\boldsymbol{x}_1-\tilde{\boldsymbol{x}}_1\right\rangle &\le L_x\|\boldsymbol{x}_1-\tilde{\boldsymbol{x}}_1\|^2+2(f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_1)).
\end{aligned}\tag{92}$$

Rearranging the above conditions, we have

$$\begin{aligned}
-2(f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_0)) &\le -\gamma\alpha(2-\gamma\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2,\\
2(f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_0)) &\le -2\xi(\delta-1)\beta\left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0),\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\right\rangle+\xi^2(\delta-1)^2\beta^2 L_y\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\|^2,\\
2(f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_1)) &\le -\delta\beta(2-\delta\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2,\\
-2(f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)-f(\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{y}}_1)) &\le -2(\gamma-1)\alpha\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1),\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\right\rangle+(\gamma-1)^2\alpha^2 L_x\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2.
\end{aligned}\tag{93}$$

Since f is convex, we have

$$2(f(\boldsymbol{x}_\star,\tilde{\boldsymbol{y}}_0)-f(\boldsymbol{x}_\star,\boldsymbol{y}_\star))\le 0,\ \ -4(f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_\star)-f(\boldsymbol{x}_\star,\boldsymbol{y}_\star))\le 0,\ \ 2(f(\boldsymbol{x}_\star,\tilde{\boldsymbol{y}}_1)-f(\boldsymbol{x}_\star,\boldsymbol{y}_\star))\le 0.\tag{94}$$

Summing up (90), (91), (93), and (94), we have

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2+\frac{2}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2+\frac{1}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2\\
&\le \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2+2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2+\left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2\\
&\quad+\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2+2\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2+\alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)\|^2\\
&\quad-\gamma\alpha(2-\gamma\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2+\xi^2(\delta-1)^2\beta^2 L_y\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\|^2\\
&\quad-\delta\beta(2-\delta\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2+(\gamma-1)^2\alpha^2 L_x\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2\\
&\quad-2(\gamma-1)\alpha\left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1),\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\right\rangle-2\xi(\delta-1)\beta\left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0),\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\right\rangle.
\end{aligned}\tag{95}$$

STEP 2. USING THE L_{xy} CONDITIONS

By definition, the Lipschitz gradient condition for L_x and L_y yields the following inequalities:

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0)\| \le L_x \|\boldsymbol{x}_0 - \boldsymbol{x}_1\|,$$
$$\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) - \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)\| \le L_y \|\boldsymbol{y}_{-1} - \boldsymbol{y}_0\|,$$

which implies

$$\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0), \boldsymbol{x}_0 - \boldsymbol{x}_1 \rangle \le L_x \|\boldsymbol{x}_0 - \boldsymbol{x}_1\|^2,$$
$$-\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) - \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \boldsymbol{y}_{-1} - \boldsymbol{y}_0 \rangle \le L_y \|\boldsymbol{y}_{-1} - \boldsymbol{y}_0\|^2,$$

or equivalently,

$$(1 - \alpha L_x) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \le \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0), \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0) \rangle ,$$
$$\xi^2 (1 - \beta L_y) \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \le \xi \langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \rangle .$$

Note that since $\xi^2 = \xi$ for both $\xi = 0$ or 1, the inequality for the $\boldsymbol{y}$ side is equivalent to

$$\xi (1 - \beta L_y) \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \le \xi \langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \rangle .$$

Therefore we can obtain the below inequalities:

$$-2(\gamma - 1)\alpha \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0), \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0) \rangle \le -2(\gamma - 1)\alpha (1 - \alpha L_x) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2,$$
$$-2\xi(\delta - 1)\beta \langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \rangle \le -2\xi(\delta - 1)\beta (1 - \beta L_y) \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2.$$

Now we can use the Lipschitz gradient condition for L_{xy} to obtain

$$\begin{aligned}
&- 2 \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0), \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) \rangle \\
&\le 2 \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0)\| \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| \\
&\le 2 L_{xy} \|\tilde{\boldsymbol{y}}_1 - \tilde{\boldsymbol{y}}_0\| \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| \\
&= 2 L_{xy} \|(\tilde{\boldsymbol{y}}_1 - \boldsymbol{y}_0) - (\tilde{\boldsymbol{y}}_0 - \boldsymbol{y}_0)\| \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| \\
&= 2 L_{xy} \|\delta\beta \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0) - \xi(\delta - 1)\beta \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| \\
&\le 2\delta\beta L_{xy} \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\| \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| + 2\xi(\delta - 1)\beta L_{xy} \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| \\
&\le \delta\beta L_{xy} \left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 \right) \\
&\quad + \xi(\delta - 1)\beta L_{xy} \left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \right),
\end{aligned}$$

where we use AM-GM for the last inequality.

Similarly, we can obtain

$$\begin{aligned}
&- 2\xi \langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0) - \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \rangle \\
&\le 2\xi \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0) - \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0)\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \\
&\le 2\xi L_{xy} \|\tilde{\boldsymbol{x}}_1 - \tilde{\boldsymbol{x}}_0\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \\
&= 2\xi L_{xy} \|(\tilde{\boldsymbol{x}}_1 - \boldsymbol{x}_0) - (\tilde{\boldsymbol{x}}_0 - \boldsymbol{x}_0)\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \\
&= 2\xi L_{xy} \|\gamma\alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) - \xi(\gamma - 1)\alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \\
&= 2\xi L_{xy} \|\gamma\alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) - (\gamma - 1)\alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \\
&\le 2\xi\gamma\alpha L_{xy} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| + 2\xi(\gamma - 1)\alpha L_{xy} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\| \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\| \\
&\le \gamma\xi\alpha L_{xy} \left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \right) \\
&\quad + (\gamma - 1)\xi\alpha L_{xy} \left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \right),
\end{aligned}$$

where the third equality is true for both $\xi = 1$ and $\xi = 0$ (where everything just becomes zero).

From this, we can deduce that

$$
\begin{aligned}
&- 2(\gamma-1)\alpha \left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1), \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) \right\rangle - 2\xi(\delta-1)\beta \left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \right\rangle \\
&= -2(\gamma-1)\alpha \left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0), \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) \right\rangle - 2\xi(\delta-1)\beta \left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \right\rangle \\
&\quad - 2(\gamma-1)\alpha \left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_0), \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0) \right\rangle \\
&\quad - 2\xi(\delta-1)\beta \left\langle \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0) - \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_0), \nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1}) \right\rangle \\
&\leq -2(\gamma-1)\alpha(1-\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 - 2\xi(\delta-1)\beta(1-\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \\
&\quad + (\gamma-1)\delta\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2\right) \\
&\quad + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2\right) \\
&\quad + \xi\gamma(\delta-1)\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2\right) \\
&\quad + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2\right).
\end{aligned}
$$

Applying this to (95), we have

$$
\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1 - \boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2 - \boldsymbol{x}_\star\|^2 \\
&\leq \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta} - \mu_y\right)\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha} - \mu_x\right)\|\boldsymbol{x}_1 - \boldsymbol{x}_\star\|^2 \\
&\quad + \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + 2\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 \\
&\quad - \gamma\alpha(2-\gamma\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi^2(\delta-1)^2\beta^2 L_y\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \\
&\quad - \delta\beta(2-\delta\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + (\gamma-1)^2\alpha^2 L_x\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad - 2(\gamma-1)\alpha(1-\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 - 2\xi(\delta-1)\beta(1-\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \\
&\quad + (\gamma-1)\delta\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2\right) \\
&\quad + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2\right) \\
&\quad + \xi\gamma(\delta-1)\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2\right) \\
&\quad + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\left(\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 + \sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2\right).
\end{aligned} \tag{96}
$$

STEP 3. SIMPLIFY USING STEP SIZE CONDITIONS

Let us gather all $\nabla_{\boldsymbol{x}}$ terms in (96), and define the sum of all such terms as

$$
\begin{aligned}
S_{\boldsymbol{x}} = {}& \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 - \gamma\alpha(2-\gamma\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&+ (\gamma-1)^2\alpha^2 L_x\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 - 2(\gamma-1)\alpha(1-\alpha L_x)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&+ (\gamma-1)\delta\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&+ \xi\gamma(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2.
\end{aligned}
$$

Rearranging terms, we have

$$
\begin{aligned}
S_{\boldsymbol{x}} &= \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad - \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 \\
&\quad + \left(2\alpha - \gamma\alpha(2-\gamma\alpha L_x) + (\gamma-1)^2\alpha^2 L_x - 2(\gamma-1)\alpha(1-\alpha L_x)\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad + \left((\gamma-1)\delta + \xi(3\gamma-2)(\delta-1)\right)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&= \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad - \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 \\
&\quad - \alpha\left(4(\gamma-1) - (2\gamma^2-1)\alpha L_x - \left((\gamma-1)\delta + \xi(3\gamma-2)(\delta-1)\right)\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}}\right) \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\leq \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad - \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2 \\
&\quad - \alpha\left(4(\gamma-1) - (2\gamma^2-1)C_1 - \left((\gamma-1)\delta + \xi(3\gamma-2)(\delta-1)\right)C_4\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2,
\end{aligned}
$$

where we use $\alpha \leq \frac{C_1}{L_x}$ and $\beta \leq \frac{C_4}{L_{xy}}\sqrt{\frac{\mu_x}{\mu_y}}$.

Since we have from (52):

$$
\begin{aligned}
4(\gamma-1) &\geq (2\gamma^2-1)C_1 + (4\gamma\delta - 3\gamma - 3\delta + 2)\,C_4 \\
&= (2\gamma^2-1)C_1 + \left((\gamma-1)\delta + (3\gamma-2)(\delta-1)\right)C_4 \\
&\geq (2\gamma^2-1)C_1 + \left((\gamma-1)\delta + \xi(3\gamma-2)(\delta-1)\right)C_4,
\end{aligned}
$$

we can deduce that

$$
-\alpha\left(4(\gamma-1) - (2\gamma^2-1)C_1 - \left((\gamma-1)\delta + \xi(3\gamma-2)(\delta-1)\right)C_4\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \leq 0
$$

and therefore

$$
\begin{aligned}
S_{\boldsymbol{x}} &\leq \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1, \tilde{\boldsymbol{y}}_1)\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad - \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0, \tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}} \cdot \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1}, \tilde{\boldsymbol{y}}_{-1})\|^2.
\end{aligned} \tag{97}
$$

Similarly, Let us gather all $\nabla_{\boldsymbol{y}}$ terms in (96), and define the sum all such terms as

$$
\begin{aligned}
S_{\boldsymbol{y}} &= 2\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \xi^2(\delta-1)^2\beta^2 L_y\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \\
&\quad - \delta\beta(2-\delta\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 - 2\xi(\delta-1)\beta(1-\beta L_y)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \\
&\quad + (\gamma-1)\delta\alpha\beta L_{xy}\sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1, \boldsymbol{y}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 \\
&\quad + \xi\gamma(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_x}{\mu_y}} \cdot \|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0, \boldsymbol{y}_{-1})\|^2.
\end{aligned}
$$

Rearranging terms, we have

$$\begin{aligned}
S_{\boldsymbol{y}} &= \left(2\beta - \delta\beta(2-\delta\beta L_y) + (\gamma-1)\delta\alpha\beta L_{xy}\sqrt{\frac{\mu_x}{\mu_y}}\right)\cdot\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2 \\
&\quad + \left(\xi^2(\delta-1)^2\beta^2 L_y - 2\xi(\delta-1)\beta(1-\beta L_y) + \xi(3\gamma-2)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_x}{\mu_y}}\right)\cdot\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\|^2 \\
&= -\beta\left(2(\delta-1) - \delta^2\beta L_y - (\gamma-1)\delta\alpha L_{xy}\sqrt{\frac{\mu_x}{\mu_y}}\right)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2 \\
&\quad - \xi(\delta-1)\beta\left(2 - (\xi(\delta-1)+2)\beta L_y - (3\gamma-2)\alpha L_{xy}\sqrt{\frac{\mu_x}{\mu_y}}\right)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\|^2 \\
&\le -\beta\left(2(\delta-1) - \delta^2 C_2 - (\gamma-1)\delta C_3\right)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2 \\
&\quad - \xi(\delta-1)\beta\left(2 - (\xi(\delta-1)+2)C_2 - (3\gamma-2)C_3\right)\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_0,\boldsymbol{y}_{-1})\|^2,
\end{aligned}$$

where we use $\beta \le \frac{C_2}{L_y}$ and $\alpha \le \frac{C_3}{L_{xy}}\sqrt{\frac{\mu_y}{\mu_x}}$.

Since we have from (53) and (54):

$$\begin{aligned}
\delta - 1 &\ge \delta^2 C_2 + (\gamma-1)\delta C_3, \\
2 &\ge (\delta+1)C_2 + (3\gamma-2)C_3 \ge (\xi(\delta-1)+2)C_2 + (3\gamma-2)C_3,
\end{aligned}$$

we can deduce that

$$S_{\boldsymbol{y}} \le -(\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2. \tag{98}$$

By (97) and (98), we can observe that (96) boils down to

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2 \\
&\le \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 + S_{\boldsymbol{x}} + S_{\boldsymbol{y}} \\
&\le \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 \\
&\quad + \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)\|^2 - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2 \\
&\quad - \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}}\cdot\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}}\cdot\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1},\tilde{\boldsymbol{y}}_{-1})\|^2 \\
&\quad - (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2,
\end{aligned}$$

or equivalently

$$\begin{aligned}
&\frac{1}{\alpha}\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 + \frac{2}{\beta}\|\boldsymbol{y}_1-\boldsymbol{y}_\star\|^2 + \frac{1}{\alpha}\|\boldsymbol{x}_2-\boldsymbol{x}_\star\|^2 \\
&\quad - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_1,\tilde{\boldsymbol{y}}_1)\|^2 + (\delta-1)\beta\|\nabla_{\boldsymbol{y}} f(\tilde{\boldsymbol{x}}_1,\boldsymbol{y}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}}\cdot\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2 \\
&\le \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_0-\boldsymbol{x}_\star\|^2 + 2\left(\frac{1}{\beta}-\mu_y\right)\|\boldsymbol{y}_0-\boldsymbol{y}_\star\|^2 + \left(\frac{1}{\alpha}-\mu_x\right)\|\boldsymbol{x}_1-\boldsymbol{x}_\star\|^2 \\
&\quad - \alpha\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0,\tilde{\boldsymbol{y}}_0)\|^2 + \xi(\gamma-1)(\delta-1)\alpha\beta L_{xy}\sqrt{\frac{\mu_y}{\mu_x}}\cdot\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{-1},\tilde{\boldsymbol{y}}_{-1})\|^2,
\end{aligned}$$

which is identical to (62) and therefore concludes the proof. □

E. Proofs used in Section 6

Here we prove all theorems related to **Alex-GDA** on bilinear problems presented in Section 6.

- In Appendix E.1 we prove Theorem 6.1 which shows the exact condition for linear convergence of **Alex-GDA** on bilinear problems.
- In Appendix E.2 we prove Theorem 6.2 which obtains iteration complexity of **Alex-GDA** for bilinear problems.
- In Appendix E.3 we prove technical propositions and lemmas used throughout the proofs in Appendix E.

E.1. Proof of Theorem 6.1

Here we prove Theorem 6.1 of Section 6, restated below for the sake of readability.

Theorem 6.1. *With a proper choice of step sizes α and β, **Alex-GDA** linearly converges to a Nash equilibrium of a bilinear problem if and only if $\gamma+\delta>2$. In this case, the exact conditions for convergent step sizes α and β are:*

$$\begin{cases}\alpha\beta<\frac{4}{(2\gamma-1)(2\delta-1)L_{xy}^2}, & \text{if } 4\gamma\delta-3(\gamma+\delta)+2\geq 0,\\ \alpha\beta<\frac{\gamma+\delta-2}{-(\gamma-1)(\delta-1)(\gamma+\delta-1)L_{xy}^2}, & \text{if } 4\gamma\delta-3(\gamma+\delta)+2<0.\end{cases}$$

Proof. For a bilinear problem $f(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{y}$, each iteration ($k\geq 0$) of **Alex-GDA** is written as $\tilde{\boldsymbol{y}}_0=\boldsymbol{y}_0$ and

$$\begin{aligned}\boldsymbol{x}_{k+1}&=\boldsymbol{x}_k-\alpha\boldsymbol{B}\tilde{\boldsymbol{y}}_k,\\ \tilde{\boldsymbol{x}}_{k+1}&=\boldsymbol{x}_k-\gamma\alpha\boldsymbol{B}\tilde{\boldsymbol{y}}_k,\\ \boldsymbol{y}_{k+1}&=\boldsymbol{y}_k+\beta\boldsymbol{B}^\top\tilde{\boldsymbol{x}}_{k+1}=\beta\boldsymbol{B}^\top\boldsymbol{x}_k+\boldsymbol{y}_k-\gamma\alpha\beta\boldsymbol{B}^\top\boldsymbol{B}\tilde{\boldsymbol{y}}_k,\\ \tilde{\boldsymbol{y}}_{k+1}&=\boldsymbol{y}_k+\delta\beta\boldsymbol{B}^\top\tilde{\boldsymbol{x}}_{k+1}=\delta\beta\boldsymbol{B}^\top\boldsymbol{x}_k+\boldsymbol{y}_k-\gamma\alpha\delta\beta\boldsymbol{B}^\top\boldsymbol{B}\tilde{\boldsymbol{y}}_k.\end{aligned}$$

This can be represented in the following matrix iteration:

$$\boldsymbol{w}_{k+1}=\begin{bmatrix}\boldsymbol{x}_{k+1}\\ \boldsymbol{y}_{k+1}\\ \tilde{\boldsymbol{y}}_{k+1}\end{bmatrix}=\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} & -\alpha\boldsymbol{B}\\ \beta\boldsymbol{B}^\top & \boldsymbol{I} & -\gamma\alpha\beta\boldsymbol{B}^\top\boldsymbol{B}\\ \delta\beta\boldsymbol{B}^\top & \boldsymbol{I} & -\gamma\alpha\delta\beta\boldsymbol{B}^\top\boldsymbol{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_k\\ \boldsymbol{y}_k\\ \tilde{\boldsymbol{y}}_k\end{bmatrix}=\boldsymbol{M}\boldsymbol{w}_k. \tag{99}$$

Consider a reduced form of singular value decomposition (SVD) of $\boldsymbol{B}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$: $\boldsymbol{U}\in\mathbb{R}^{d_x\times s},\boldsymbol{V}\in\mathbb{R}^{d_y\times s},\boldsymbol{\Sigma}\in\mathbb{R}^{s\times s}$ where $s=\operatorname{rank}(\boldsymbol{B})$. Note that $\boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}$, $\boldsymbol{V}^\top\boldsymbol{V}=\boldsymbol{I}$, and $\boldsymbol{\Sigma}=\operatorname{diag}(\sigma_1,\dots,\sigma_s)$ is a diagonal matrix with non-zero diagonal entries ($0<\mu_{xy}\leq\sigma_i\leq L_{xy}$ for all $i=1,\dots,s$). Then the power of the matrix $\boldsymbol{M}$ defined in Equation (99) can be decomposed as follows for $k\geq 1$.

$$\boldsymbol{M}^k=\underbrace{\begin{bmatrix}\boldsymbol{U} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{V} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{V}\end{bmatrix}}_{=:\boldsymbol{W}}\underbrace{\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} & -\alpha\boldsymbol{\Sigma}\\ \beta\boldsymbol{\Sigma} & \boldsymbol{I} & -\gamma\alpha\beta\boldsymbol{\Sigma}^2\\ \delta\beta\boldsymbol{\Sigma} & \boldsymbol{I} & -\gamma\alpha\delta\beta\boldsymbol{\Sigma}^2\end{bmatrix}^k}_{=:\widetilde{\boldsymbol{M}}^k}\begin{bmatrix}\boldsymbol{U}^\top & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{V}^\top & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{V}^\top\end{bmatrix}+\begin{bmatrix}\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^\top & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^\top & \boldsymbol{0}\end{bmatrix}$$

From this matrix decomposition, a decomposition of the ambient space $\mathbb{R}^{d_x+d_y+d_y}$ naturally arises: a space $\mathcal{N}=\operatorname{null}(\boldsymbol{B})\times\operatorname{null}(\boldsymbol{B}^\top)\times\operatorname{null}(\boldsymbol{B}^\top)$ and its orthogonal complement $\mathcal{N}^\perp=\operatorname{row}(\boldsymbol{B})\times\operatorname{row}(\boldsymbol{B}^\top)\times\operatorname{row}(\boldsymbol{B}^\top)$. The $\mathcal{N}$-component of the iterate $\boldsymbol{w}_k$ is always fixed as

$$\begin{bmatrix}(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_0\\ (\boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{y}_0\\ (\boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{y}_0\end{bmatrix}$$

and does not move at all, while the $\mathcal{N}^\perp$-component of $\boldsymbol{w}_k$ belong to $\mathcal{N}^\perp$ even after each iteration. Since $\operatorname{null}(\boldsymbol{B})\times\operatorname{null}(\boldsymbol{B}^\top)$ is the space of all Nash equilibria of the bilinear problem, now it is enough to show that the $\mathcal{N}^\perp$-component converges to the origin; as a result, the iterates $(\boldsymbol{x}_k,\boldsymbol{y}_k)$ converge to a Nash equilibrium

$$\boldsymbol{z}_\star:=((\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}_0,(\boldsymbol{I}-\boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{y}_0). \tag{100}$$

To this end, we may assume that the initial iterate $\boldsymbol{w}_0$ belongs to $\mathcal{N}^\perp$ from now on. Then, by reasoning above, every iterate $\boldsymbol{w}_k$ belongs to $\mathcal{N}^\perp$ and satisfies

$$\boldsymbol{w}_k = \boldsymbol{W}\widetilde{\boldsymbol{M}}^k\boldsymbol{W}^\top\boldsymbol{w}_0. \tag{101}$$

We first claim that it suffices to show $\rho(\widetilde{\boldsymbol{M}}) < 1$ to obtain (the necessary and sufficient condition for) the convergence $\boldsymbol{w}_k \to \boldsymbol{0}$. To prove the claim, let $\widetilde{\boldsymbol{w}}_k := \boldsymbol{W}^\top\boldsymbol{w}_k$. Then we have $\widetilde{\boldsymbol{w}}_k = \widetilde{\boldsymbol{M}}^k\widetilde{\boldsymbol{w}}_0$. By applying the theory of matrix iteration (Proposition B.4), $\rho(\widetilde{\boldsymbol{M}}) < 1$ if and only if $\widetilde{\boldsymbol{w}}_k \to \boldsymbol{0}$. Moreover, since $\boldsymbol{w}_k \in \mathcal{N}^\perp$, $\boldsymbol{W}\widetilde{\boldsymbol{w}}_k = \boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{w}_k = \boldsymbol{w}_k$, and thus $\widetilde{\boldsymbol{w}}_k \to \boldsymbol{0}$ if and only if $\boldsymbol{w}_k \to \boldsymbol{0}$. Therefore, the rest of the proof is dedicated to finding the condition for $\rho(\widetilde{\boldsymbol{M}}) < 1$.

Note that the matrix $\widetilde{\boldsymbol{M}} \in \mathbb{R}^{3s\times 3s}$ does not have 1 as an eigenvalue. If it does, there exist vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^s$ such that $\widetilde{\boldsymbol{M}}\begin{bmatrix}\boldsymbol{a}^\top & \boldsymbol{b}^\top & \boldsymbol{c}^\top\end{bmatrix}^\top = \begin{bmatrix}\boldsymbol{a}^\top & \boldsymbol{b}^\top & \boldsymbol{c}^\top\end{bmatrix}^\top$. It implies that

$$\begin{aligned}\boldsymbol{a} - \alpha\boldsymbol{\Sigma}\boldsymbol{c} &= \boldsymbol{a},\\ \beta\boldsymbol{\Sigma}\boldsymbol{a} + \boldsymbol{b} - \gamma\alpha\beta\boldsymbol{\Sigma}^2\boldsymbol{c} &= \boldsymbol{b},\\ \delta\beta\boldsymbol{\Sigma}\boldsymbol{a} + \boldsymbol{b} - \gamma\alpha\delta\beta\boldsymbol{\Sigma}^2\boldsymbol{c} &= \boldsymbol{c},\end{aligned}$$

which implies that $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{c} = 0$ because $\boldsymbol{\Sigma}$ is nonsingular. Thus, 1 cannot have an associated nonzero eigenvector of $\boldsymbol{M}$.

To inspect the eigenvalues of $\widetilde{\boldsymbol{M}}$, we now apply the theory of Schur complement (Haynsworth, 1968; Zhang, 2006): namely, $\det\left(\begin{bmatrix}A & B\\ C & D\end{bmatrix}\right) = \det(A)\det(D - CA^{-1}B)$. Writing the characteristic polynomial of $\widetilde{\boldsymbol{M}}$,

$$\begin{aligned}\det(\lambda\boldsymbol{I} - \widetilde{\boldsymbol{M}}) &= \det\left(\begin{bmatrix}(\lambda-1)\boldsymbol{I} & \boldsymbol{0} & \alpha\boldsymbol{\Sigma}\\ -\beta\boldsymbol{\Sigma} & (\lambda-1)\boldsymbol{I} & \gamma\alpha\beta\boldsymbol{\Sigma}^2\\ -\delta\beta\boldsymbol{\Sigma} & -\boldsymbol{I} & \lambda\boldsymbol{I} + \gamma\alpha\delta\beta\boldsymbol{\Sigma}^2\end{bmatrix}\right)\\ &= \det\left(\begin{bmatrix}(\lambda-1)\boldsymbol{I} & \boldsymbol{0}\\ -\beta\boldsymbol{\Sigma} & (\lambda-1)\boldsymbol{I}\end{bmatrix}\right)\det\left(\lambda\boldsymbol{I} + \gamma\alpha\delta\beta\boldsymbol{\Sigma}^2 + \frac{\alpha}{\lambda-1}\begin{bmatrix}\delta\beta\boldsymbol{\Sigma} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ -\frac{\beta}{\lambda-1}\boldsymbol{\Sigma} & \boldsymbol{I}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\Sigma}\\ \gamma\beta\boldsymbol{\Sigma}^2\end{bmatrix}\right)\\ &= (\lambda-1)^{2s}\det\left(\lambda\boldsymbol{I} + \gamma\alpha\delta\beta\boldsymbol{\Sigma}^2 + \frac{\alpha}{\lambda-1}\begin{bmatrix}\delta\beta\boldsymbol{\Sigma} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \frac{\beta}{\lambda-1}\boldsymbol{\Sigma} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}\\ \gamma\beta\boldsymbol{\Sigma}^2\end{bmatrix}\right)\\ &= \det\left((\lambda-1)^2\boldsymbol{I} + \alpha\beta\left(\gamma(\lambda-1)+1\right)\left(\delta(\lambda-1)+1\right)\boldsymbol{\Sigma}^2\right) = 0.\end{aligned}$$

Hence, for each eigenvalue σ_i^2 of $\boldsymbol{\Sigma}^2$, the roots λ of a cubic polynomial

$$\begin{aligned}P_i(\lambda) &:= \lambda(\lambda-1)^2 + \alpha\beta\sigma_i^2\left(\gamma(\lambda-1)+1\right)\left(\delta(\lambda-1)+1\right)\\ &= \lambda^3 - (2-\phi_i\gamma\delta)\lambda^2 + \left\{1 - \phi_i(2\gamma\delta - \gamma - \delta)\right\}\lambda + \phi_i(\gamma-1)(\delta-1)\end{aligned} \tag{102}$$

are eigenvalues of $\widetilde{\boldsymbol{M}}$, where $\phi_i := \alpha\beta\sigma_i^2 > 0$. To obtain a necessary and sufficient condition of $|\lambda| < 1$, we apply Proposition D.3:

$$\phi_i\gamma\delta - 2 + \phi_i(\gamma-1)(\delta-1) < 2 - \phi_i(2\gamma\delta - \gamma - \delta), \tag{103}$$
$$\phi_i\gamma\delta - 2 + \phi_i(\gamma-1)(\delta-1) > -2 + \phi_i(2\gamma\delta - \gamma - \delta), \tag{104}$$
$$\phi_i\gamma\delta - 2 - 3\phi_i(\gamma-1)(\delta-1) < 2 + \phi_i(2\gamma\delta - \gamma - \delta), \tag{105}$$
$$\phi_i\gamma\delta - 2 - 3\phi_i(\gamma-1)(\delta-1) > -2 - \phi_i(2\gamma\delta - \gamma - \delta), \tag{106}$$
$$\phi_i(\gamma-1)(\delta-1)(\phi_i(\gamma-1)(\delta-1) - \phi_i\gamma\delta + 2) + 1 - \phi_i(2\gamma\delta - \gamma - \delta) < 1, \tag{107}$$

which are equivalent to

$$\phi_i \overset{(104)}{>} 0, \quad \text{(which is already true,)} \tag{108}$$

$$\gamma + \delta \overset{(106)}{>} \frac{3}{2}, \tag{109}$$

$$\phi_i(2\gamma - 1)(2\delta - 1) \overset{(103)}{<} 4, \tag{110}$$

$$\phi_i(1 - 4(\gamma - 1)(\delta - 1)) \overset{(105)}{<} 4, \tag{111}$$

$$-(\gamma - 1)(\delta - 1)(\gamma + \delta - 1)\phi_i \overset{(107)}{<} \gamma + \delta - 2. \tag{112}$$

To make these conditions more concise and interpretable, we conduct a case analysis on γ and δ to know which condition among them is essential for having $|\lambda| < 1$ (in fact, $\gamma + \delta > \frac{3}{2}$ is not enough yet!) and to identify what condition on ϕ_i should suffice for each case.

Case 1. $(\gamma-1)(\delta-1) \geq 0$ **and** $\gamma+\delta > 2$**.** Note that Equation (112) is true. Also, $(2\gamma-1)(2\delta-1) > 1-4(\gamma-1)(\delta-1)$ since

$$\begin{aligned}(2\gamma - 1)(2\delta - 1) - 1 + 4(\gamma - 1)(\delta - 1) &= 2(4\gamma\delta - 3(\gamma + \delta) + 2) \\ &= 8(\gamma - 1)(\delta - 1) + 2(\gamma + \delta - 2) > 0.\end{aligned}$$

Thus, Equation (110) implies Equation (111). It means that Equation (110) alone is enough: $|\lambda| < 1$ if

$$\phi_i < \frac{4}{(2\gamma - 1)(2\delta - 1)}.$$

Case 2. $(\gamma - 1)(\delta - 1) \geq 0$ **and** $\frac{3}{2} < \gamma + \delta \leq 2$**.** For this case, it is impossible to satisfy all four conditions (109)–(112) at the same time. We prove it by contradiction. Note that $0 < (2\gamma - 1)(2\delta - 1) < 1 - 4(\gamma - 1)(\delta - 1)$ since

$$\begin{aligned}(2\gamma - 1)(2\delta - 1) &= 4(\gamma - 1)(\delta - 1) + 2(\gamma + \delta) - 3 > 0, \\ (2\gamma - 1)(2\delta - 1) - 1 + 4(\gamma - 1)(\delta - 1) &= 2(4\gamma\delta - 3(\gamma + \delta) + 2) \\ &\leq 2((\gamma + \delta)^2 - 3(\gamma + \delta) + 2) \\ &= 2(\gamma + \delta - 1)(\gamma + \delta - 2) < 0.\end{aligned}$$

From Equation (111) and Equation (112), it must hold that

$$\frac{2 - \gamma - \delta}{(\gamma - 1)(\delta - 1)(\gamma + \delta - 1)} < \phi_i < \frac{4}{1 - 4(\gamma - 1)(\delta - 1)}.$$

However, it implies that

$$\begin{aligned}&4(\gamma - 1)(\delta - 1)(\gamma + \delta - 1) - (2 - \gamma - \delta)(1 - 4(\gamma - 1)(\delta - 1)) \\ &= 4\gamma\delta - 3(\gamma + \delta) + 2 > 0,\end{aligned}$$

which is a contradiction.

Case 3. $(\gamma - 1)(\delta - 1) < 0$ **and** $\frac{3}{2} < \gamma + \delta \leq 2$**.** This case is also impossible since it contradicts Equation (112).

Case 4. $(\gamma - 1)(\delta - 1) < 0$**,** $\gamma + \delta > 2$**, and** $4\gamma\delta - 3(\gamma + \delta) + 2 \geq 0$**.** In this case, it holds that

$$0 < \frac{-(\gamma - 1)(\delta - 1)(\gamma + \delta - 1)}{\gamma + \delta - 2} \leq \frac{1 - 4(\gamma - 1)(\delta - 1)}{4} \leq \frac{(2\gamma - 1)(2\delta - 1)}{4}$$

since

$$(2\gamma - 1)(2\delta - 1) - 1 + 4(\gamma - 1)(\delta - 1) = 2(4\gamma\delta - 3(\gamma + \delta) + 2) \geq 0$$

and

$$\begin{aligned}&4(\gamma-1)(\delta-1)(\gamma+\delta-1)+(\gamma+\delta-2)(1-4(\gamma-1)(\delta-1))\\&=4\gamma\delta-3(\gamma+\delta)+2\geq 0.\end{aligned}$$

Thus, Equation (110) implies Equation (111) and Equation (112). Since the rightmost term is positive, we have $|\lambda|<1$ if

$$\phi_i<\frac{4}{(2\gamma-1)(2\delta-1)}.$$

Case 5. $(\gamma-1)(\delta-1)<0$**,** $\gamma+\delta>2$**, and** $4\gamma\delta-3(\gamma+\delta)+2<0$**.** In this case, it holds that

$$\frac{(2\gamma-1)(2\delta-1)}{4}<\frac{1-4(\gamma-1)(\delta-1)}{4}<\frac{-(\gamma-1)(\delta-1)(\gamma+\delta-1)}{\gamma+\delta-2}$$

Thus, Equation (112) implies Equation (110) and Equation (111). Since the rightmost term is positive, we have $|\lambda|<1$ if

$$\phi_i<\frac{\gamma+\delta-2}{-(\gamma-1)(\delta-1)(\gamma+\delta-1)}.$$

Combining all these five cases,

1. (**Case 2 + Case 3**) If $\gamma+\delta\leq 2$, the polynomial $P_i(\lambda)$ must have a root outside of the open unit disk; hence, the matrix iteration in Equation (101) diverges.

2. (**Case 1 + Case 4**) If $\gamma+\delta>2$ and $4\gamma\delta-3(\gamma+\delta)+2\geq 0$ (which includes the case of $\gamma+\delta>2$, $\gamma\geq 1$, and $\delta\geq 1$), all the roots of the polynomial $P_i(\lambda)$ lie on the open unit disk $|\lambda|<1$ if

   $$\phi_i<\frac{4}{(2\gamma-1)(2\delta-1)}.$$

   Hence if we choose step sizes α and β such that

   $$\alpha\beta<\frac{4}{(2\gamma-1)(2\delta-1)L_{xy}^2},$$

   then all the eigenvalues of $\widetilde{\boldsymbol{M}}$ lie on the open unit disk; the matrix iteration in Equation (101) does converge.

3. (**Case 5**) If $\gamma+\delta>2$ and $4\gamma\delta-3(\gamma+\delta)+2<0$, all the roots of the polynomial $P_i(\lambda)$ lie on the open unit disk $|\lambda|<1$ if

   $$\phi_i<\frac{\gamma+\delta-2}{-(\gamma-1)(\delta-1)(\gamma+\delta-1)}.$$

   Hence if we choose step sizes α and β such that

   $$\alpha\beta<\frac{\gamma+\delta-2}{-(\gamma-1)(\delta-1)(\gamma+\delta-1)L_{xy}^2},$$

   then all the eigenvalues of $\widetilde{\boldsymbol{M}}$ lie on the open unit disk; the matrix iteration in Equation (101) does converge.

This proves the theorem.

□

E.2. Proof of Theorem 6.2

Here we prove Theorem 6.2 of Section 6, restated below for the sake of readability.

Theorem 6.2. *For general $\gamma \geq 1$ and $\delta \geq 1$ such that $\gamma + \delta > 2$, If we choose the step sizes α and β so that $\alpha\beta = \frac{1}{C_{\gamma,\delta} L_{xy}^2}$ where $C_{\gamma,\delta} > 0$ is a constant that only depends on γ and δ, an iteration complexity upper bound of **Alex-GDA** is*

$$\mathcal{O}\left(\frac{C_{\gamma,\delta}}{\gamma+\delta-2} \cdot \left(\frac{L_{xy}}{\mu_{xy}}\right)^2 \cdot \log\left(\frac{\|\boldsymbol{w}_0\|^2}{\epsilon}\right)\right),$$

where $\|\boldsymbol{w}_0\|^2 = \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2 + 2\|\boldsymbol{y}_0 - \boldsymbol{y}_\star\|^2$ and $\boldsymbol{z}_\star = (\boldsymbol{x}_\star, \boldsymbol{y}_\star)$ is a uniquely determined Nash equilibrium if $\boldsymbol{z}_0$ is given.

*If $\delta = 1$, the optimal rate exponent of **Alex-GDA** is*

$$\lim_{k\to\infty} \frac{\|\boldsymbol{z}_k - \boldsymbol{z}_\star\|}{\|\boldsymbol{z}_{k-1} - \boldsymbol{z}_\star\|} = \sqrt{\frac{L_{xy}^2 - \mu_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}},$$

where the optimal choice of parameters are

$$\alpha\beta = \frac{2\mu_{xy}^2/L_{xy}^2}{L_{xy}^2 + \mu_{xy}^2}, \quad \gamma = 1 + \frac{L_{xy}^2}{\mu_{xy}^2}.$$

Proof. Recall that the Nash equilibrium that the iterates converges to is already characterized in Equation (100). So, as in the proof in Appendix E.2, we again assume that $\boldsymbol{w}_0$ (defined in Equation (99)) belongs to $\mathcal{N}^\perp = \mathrm{row}(\boldsymbol{B}) \times \mathrm{row}(\boldsymbol{B}^\top) \times \mathrm{row}(\boldsymbol{B}^\top)$ and we inspect the convergence (to $\mathbf{0}$) of the sequence (101). For this reason, we analyze the spectral radius of the matrix $\widetilde{\boldsymbol{M}}$ (defined in Equation (99)). This will directly give us a convergence rate as well as iteration complexity ($\tilde{\mathcal{O}}\left(\frac{1}{1-\rho(\widetilde{\boldsymbol{M}})}\right)$).

We divide the proof into two parts: the case of general parameters $\gamma \geq 1$ and $\delta \geq 1$, and the case of $\delta = 1$. Throughout the proof, we keep the notation consistent with the proof of Theorem 6.1 in Appendix E.1.

E.2.1. The Case of General $\gamma \geq 1$ and $\delta \geq 1$

We have to find an upper bound of $|\lambda|$ which is strictly smaller than 1, whose difference with 1 is not negligible. Hence, we use a slightly smaller bound $\phi_i \leq \frac{2}{(2\gamma-1)(2\delta-1)}$ than that in Theorem 6.1.

With some substitutions

$$\psi_i := \alpha\beta\gamma\delta\sigma_i^2 = \gamma\delta\phi_i > 0, \quad \Gamma := 1 - \frac{1}{\gamma} \in [0,1), \quad \Delta := 1 - \frac{1}{\delta} \in [0,1),$$

we can rewrite the polynomial $P_i(\lambda)$ as

$$\begin{aligned} P_i(\lambda) &= \lambda(\lambda-1)^2 + \psi_i(\lambda-\Gamma)(\lambda-\Delta) \\ &= \lambda^3 - (2-\psi_i)\lambda^2 + \{1 - \psi_i(\Gamma+\Delta)\}\lambda + \psi_i\Gamma\Delta. \end{aligned} \tag{113}$$

Since $P_i(0) = \psi_i\Gamma\Delta \geq 0$ and

$$P_i\left(-\frac{1}{2}\right) = -\frac{9}{8} + \psi_i\left(\Gamma + \frac{1}{2}\right)\left(\Delta + \frac{1}{2}\right) < 0$$

holds because

$$\begin{aligned} \psi_i &\leq \frac{2}{(\Gamma+1)(\Delta+1)} = \frac{2\gamma\delta}{(2\gamma-1)(2\delta-1)} = \frac{9\gamma\delta}{2\left(3\gamma-\frac{3}{2}\right)\left(3\delta-\frac{3}{2}\right)} \\ &< \frac{9\gamma\delta}{2\left(3\gamma-2\right)\left(3\delta-2\right)} = \frac{9}{8\left(\Gamma+\frac{1}{2}\right)\left(\Delta+\frac{1}{2}\right)}. \end{aligned}$$

Thus, there exists a non-positive real root $-r \in (-\frac{1}{2}, 0]$.

We can show that there is no positive real root if ψ_i is small enough.

Proposition E.1. *The polynomial $P_i(\lambda)$ defined in Equation (113) has no positive real root if*

$$\psi_i|\Gamma-\Delta| \le \min\left\{(1-\Gamma)^2,(1-\Delta)^2\right\}.$$

The proof of this proposition can be found in Appendix E.3.1. From the root coefficient relationship, we know that the sum of three roots of $P_i(\lambda)$ equals $2-\psi_i>0$, which holds because $\psi_i \le \frac{2}{(\Gamma+1)(\Delta+1)} < 2$. Hence, $P_i(\lambda)$ must have a single real root $-r \le 0$ and two complex conjugate roots c and $\bar{c}$, where $\Re[c]>0$.

Note that we have another bound for the unique real root. Plugging in $\lambda=-r$ to $P_i(\lambda)=0$, we have

$$\begin{aligned}\{1-\psi_i(\Gamma+\Delta)\}(-r)+\psi_i\Gamma\Delta = r^3+(2-\psi_i)r^2 \ge 0,\\ \therefore r \le \frac{\psi_i\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}.\end{aligned} \tag{114}$$

Again from the root coefficient relationship, we know that

$$\begin{aligned}-r+2\Re[c] &= 2-\psi_i,\\ -2r\Re[c]+|c|^2 &= 1-\psi_i(\Gamma+\Delta).\end{aligned}$$

Plugging one into another, we have an expression of the squared absolute value of a complex root in terms of r as

$$\begin{aligned}|c|^2 &= 1-\psi_i(\Gamma+\Delta)+r(2-\psi_i+r)\\ &\overset{(114)}{\le} 1-\psi_i(\Gamma+\Delta)+\frac{\psi_i\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\left(2-\psi_i+\frac{\psi_i\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\right)\\ &= 1-\psi_i\left\{\Gamma+\Delta-\frac{\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\left(2-\psi_i+\frac{\psi_i\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\right)\right\}\end{aligned} \tag{115}$$

To show that $|c|^2$ is strictly smaller than 1, we want to show that

$$\Gamma+\Delta-\frac{\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\left(2-\psi_i+\frac{\psi_i\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\right)>0.$$

In fact, this is shown in the following proposition.

Proposition E.2.

$$\Gamma+\Delta-\frac{\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\left(2-\psi_i+\frac{\psi_i\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)}\right) \ge \frac{1}{4}(\Gamma+\Delta-2\Gamma\Delta)>0$$

if $\psi_i \le \frac{\Gamma+\Delta-2\Gamma\Delta}{2(\Gamma+\Delta)^2}$.

The proof of this proposition can be found in Appendix E.3.2. Therefore, gathering the fact that $|-r|^2<\frac{1}{4}$, Equation (115), and Proposition E.2, for every root λ of the polynomial $P_i(\lambda)$,

$$\begin{aligned}|\lambda|^2 &< \max\left\{\frac{1}{4}, 1-\frac{\psi_i}{4}(\Gamma+\Delta-2\Gamma\Delta)\right\}\\ &= \max\left\{\frac{1}{4}, 1-\frac{1}{4}\alpha\beta\sigma_i^2(\gamma+\delta-2)\right\},\end{aligned} \tag{116}$$

where

$$\psi_i \le \min\left\{\frac{2}{(\Gamma+1)(\Delta+1)}, \frac{\min\left\{(1-\Gamma)^2,(1-\Delta)^2\right\}}{|\Gamma-\Delta|}, \frac{\Gamma+\Delta-2\Gamma\Delta}{2(\Gamma+\Delta)^2}\right\},$$

or equivalently,

$$\alpha\beta\sigma_i^2 \le \frac{1}{C_{\gamma,\delta}},$$
$$\text{where } C_{\gamma,\delta} := \max\left\{\frac{(2\gamma-1)(2\delta-1)}{2}, |\gamma-\delta|\max\{\gamma,\delta\}^2, \frac{2(2\gamma\delta-\gamma-\delta)^2}{\gamma+\delta-2}\right\}. \tag{117}$$

Hence, if we choose step sizes α and β such that $\alpha\beta = \frac{1}{C_{\gamma,\delta}L_{xy}^2}$, the bound in Equation (116) holds for all $i = 1, ..., s$, thereby we obtain a strict upper bound of spectral radius of the matrix $\widetilde{\boldsymbol{M}}$ as follows:

$$\begin{aligned}\rho(\widetilde{\boldsymbol{M}})^2 &< \max\left\{\frac{1}{4}, 1-\frac{1}{4}\alpha\beta\mu_{xy}^2(\gamma+\delta-2)\right\},\\ &= \max\left\{\frac{1}{4}, 1-\frac{\gamma+\delta-2}{4C_{\gamma,\delta}}\frac{\mu_{xy}^2}{L_{xy}^2}\right\}.\end{aligned}$$

In conclusion, the matrix iteration in Equation (101) can satisfy $\|\boldsymbol{w}_k\|^2 < \epsilon$ with

$$k = \mathcal{O}\left(\max\left\{1, \frac{C_{\gamma,\delta}}{\gamma+\delta-2}\cdot\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\left(\frac{\|\boldsymbol{w}_0\|^2}{\epsilon}\right)\right)$$

iterations.

Remark E.3. One may notice that the constant $C_{\gamma,\delta}$ defined in Equation (117) may grow as γ^3 or δ^3, which can make the range of step size with certified convergence rate shrink and degrade the iteration complexity. However, when $\delta = 1$, our analysis gets simpler and we can choose an optimal set of parameters α, β, and γ to "optimize" the spectral radius (and thus the convergence rate).

E.2.2. The case of $\delta = 1$ ($\gamma > 1$)

Let us go back to the polynomial $P_i(\lambda)$ (Equation (102)). If $\delta = 1$ (and thus we choose $\gamma > 1$), the polynomial becomes

$$P_i^{(\delta=1)}(\lambda) = \lambda\left\{(\lambda-1)^2 + \alpha\beta\sigma_i^2(\gamma\lambda-(\gamma-1))\right\}.$$

So, we know one root exactly: $\lambda = 0$. Since we want a small absolute value of eigenvalues but 0 is a trivial lower bound of $|\lambda|$, we only have to care about the other two roots: $(\lambda-1)^2 + \alpha\beta\sigma_i^2(\gamma\lambda-(\gamma-1)) = 0$, or

$$\begin{aligned}\lambda_0 &:= 1 - \frac{\gamma\alpha\beta\sigma_i^2 - \sqrt{(2-\gamma\alpha\beta\sigma_i^2)^2 - 4(1-(\gamma-1)\alpha\beta\sigma_i^2)}}{2},\\ \lambda_1 &:= 1 - \frac{\gamma\alpha\beta\sigma_i^2 + \sqrt{(2-\gamma\alpha\beta\sigma_i^2)^2 - 4(1-(\gamma-1)\alpha\beta\sigma_i^2)}}{2}.\end{aligned}$$

The maximum absolute value of eigenvalues can be calculated as

$$\begin{aligned}&\max\{|\lambda_0|, |\lambda_1|\}\\ &= \begin{cases}\sqrt{1-(\gamma-1)\alpha\beta\sigma_i^2} & \text{if } (2-\gamma\alpha\beta\sigma_i^2)^2 \le 4(1-(\gamma-1)\alpha\beta\sigma_i^2),\\ \left|1-\frac{\gamma\alpha\beta\sigma_i^2}{2}\right| + \frac{\sqrt{(2-\gamma\alpha\beta\sigma_i^2)^2 - 4(1-(\gamma-1)\alpha\beta\sigma_i^2)}}{2} & \text{if } (2-\gamma\alpha\beta\sigma_i^2)^2 > 4(1-(\gamma-1)\alpha\beta\sigma_i^2).\end{cases}\\ &= \begin{cases}\sqrt{1-(\gamma-1)\alpha\beta\sigma_i^2} & \text{if } \gamma^2\alpha\beta\sigma_i^2 \le 4,\\ \left|1-\frac{\gamma\alpha\beta\sigma_i^2}{2}\right| + \frac{\sqrt{(\gamma^2\alpha\beta\sigma_i^2-4)\alpha\beta\sigma_i^2}}{2} & \text{if } \gamma^2\alpha\beta\sigma_i^2 > 4.\end{cases}\\ &=: r(\alpha,\beta,\gamma,\sigma_i^2)\end{aligned} \tag{118}$$

Thus, if we want to optimize the spectral radius $\rho(\widetilde{M})$ (which directly gives the convergence rate exponent) by choosing parameters α, β, and γ, we have to solve the following minimax problem:

$$\min_{\alpha,\beta,\gamma} \max_{i=1,\dots,s} r(\alpha,\beta,\gamma,\sigma_i^2).$$

Suppose $L_{xy} = \sigma_1 \geq \cdots \geq \sigma_s = \mu_{xy}$. We consider 3 cases:

Case 1. $\gamma^2\alpha\beta L_{xy}^2 \leq 4$**.** In this case, $\gamma^2\alpha\beta\sigma_i^2 \leq 4$ holds for all $i = 1,\dots,s$, and then $r(\alpha,\beta,\gamma,\sigma_i^2) = \sqrt{1-(\gamma-1)\alpha\beta\sigma_i^2}$ is a decreasing function of σ_i^2. Hence, it suffices to minimize $\sqrt{1-(\gamma-1)\alpha\beta\mu_{xy}^2}$ over α, β, and γ. The optimal choice of $\alpha\beta$ is $\frac{4}{\gamma^2 L_{xy}^2}$ which comes from the condition $\gamma^2\alpha\beta L_{xy}^2 \leq 4$, so we minimize $\sqrt{1-\frac{4(\gamma-1)\mu_{xy}^2}{\gamma^2 L_{xy}^2}}$ over γ. The optimal γ is 2, so the optimal spectral radius is $\sqrt{1-\frac{\mu_{xy}^2}{L_{xy}^2}}$, which can be obtained with $\alpha\beta = \frac{1}{L_{xy}^2}$ and $\gamma = 2$.

Case 2. $\gamma^2\alpha\beta\mu_{xy}^2 \geq 4$**.** Note that

$$\left|1-\frac{\gamma\alpha\beta\sigma_i^2}{2}\right| + \frac{\sqrt{(\gamma^2\alpha\beta\sigma_i^2-4)\alpha\beta\sigma_i^2}}{2}$$

is an increasing function in terms of $\sigma_i^2 \geq \frac{4}{\gamma^2\alpha\beta}$. This can be shown by proving that

$$1-\frac{\gamma\alpha\beta\sigma_i^2}{2} + \frac{\sqrt{(\gamma^2\alpha\beta\sigma_i^2-4)\alpha\beta\sigma_i^2}}{2}$$

and

$$-1+\frac{\gamma\alpha\beta\sigma_i^2}{2} + \frac{\sqrt{(\gamma^2\alpha\beta\sigma_i^2-4)\alpha\beta\sigma_i^2}}{2}$$

are both increasing functions in terms of σ_i^2. The latter case is easy, so we show for the former one: using the derivative in σ_i^2,

$$\frac{d}{d\sigma_i^2}\left(-\gamma\alpha\beta\sigma_i^2 + \sqrt{(\gamma^2\alpha\beta\sigma_i^2-4)\alpha\beta\sigma_i^2}\right) = -\gamma\alpha\beta + \frac{\gamma^2\alpha^2\beta^2\sigma_i^2 - 2\alpha\beta}{\sqrt{\gamma^2\alpha^2\beta^2\sigma_i^4 - 4\alpha\beta\sigma_i^2}} > 0,$$

So it suffices to minimize

$$\left|1-\frac{\gamma\alpha\beta L_{xy}^2}{2}\right| + \frac{\sqrt{(\gamma^2\alpha\beta L_{xy}^2-4)\alpha\beta L_{xy}^2}}{2}$$

over α, β, and γ. In fact, this is also an increasing function in terms of $\alpha\beta \geq \frac{4}{\gamma^2\mu_{xy}^2}$ and the optimal choice of $\alpha\beta$ is $\frac{4}{\gamma^2\mu_{xy}^2}$ (which comes from the condition $\gamma^2\alpha\beta\mu_{xy}^2 \geq 4$). So it is left to minimize

$$\left|1-\frac{2L_{xy}^2}{\gamma\mu_{xy}^2}\right| + \frac{2L_{xy}}{\gamma\mu_{xy}}\sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}-1}$$

over γ. This has a minimum $\sqrt{1-\frac{\mu_{xy}^2}{L_{xy}^2}}$ at $\gamma = \frac{2L_{xy}^2}{\mu_{xy}^2}$. Hence, the optimal spectral radius is $\sqrt{1-\frac{\mu_{xy}^2}{L_{xy}^2}}$, which is achieved with $\gamma = \frac{2L_{xy}^2}{\mu_{xy}^2}$ and $\alpha\beta = \frac{\mu_{xy}^2}{L_{xy}^4}$.

Case 3. $\gamma^2\alpha\beta\mu_{xy}^2 \le 4 \le \gamma^2\alpha\beta L_{xy}^2$**.** Maximizing $r(\alpha,\beta,\gamma,\sigma_i^2)$ over $i=1,\dots,s$, we only need to obtain

$$\min_{\alpha,\beta,\gamma}\max\left\{\sqrt{1-(\gamma-1)\alpha\beta\mu_{xy}^2},\left|1-\frac{\gamma\alpha\beta L_{xy}^2}{2}\right|+\frac{\sqrt{(\gamma^2\alpha\beta L_{xy}^2-4)\alpha\beta L_{xy}^2}}{2}\right\}.$$

For a fixed γ, the optimal $X:=\alpha\beta$ is uniquely attained when

$$\sqrt{1-(\gamma-1)\mu_{xy}^2X}=\left|1-\frac{\gamma L_{xy}^2X}{2}\right|+\frac{\sqrt{\gamma^2L_{xy}^4X^2-4L_{xy}^2X}}{2}, \tag{119}$$

because the left hand side decreases in X but the right hand side increases in X, as well as

$$\sqrt{1-(\gamma-1)\mu_{xy}^2\frac{4}{\gamma^2L_{xy}^2}}\ge\left|1-\frac{\gamma L_{xy}^2}{2}\frac{4}{\gamma^2L_{xy}^2}\right|+0 \tag{120}$$

and

$$\sqrt{1-(\gamma-1)\mu_{xy}^2\frac{4}{\gamma^2\mu_{xy}^2}}\le\left|1-\frac{\gamma L_{xy}^2}{2}\frac{4}{\gamma^2\mu_{xy}^2}\right|+\frac{\sqrt{\gamma^2L_{xy}^4\left(\frac{4}{\gamma^2\mu_{xy}^2}\right)^2-4L_{xy}^2\left(\frac{4}{\gamma^2\mu_{xy}^2}\right)}}{2}. \tag{121}$$

Equation (120) can be shown as

$$\left(1-(\gamma-1)\mu_{xy}^2\frac{4}{\gamma^2L_{xy}^2}\right)-\left(1-\frac{2}{\gamma}\right)^2=\frac{4(\gamma-1)}{\gamma^2}\left(1-\frac{\mu_{xy}^2}{L_{xy}^2}\right)\ge 0.$$

In addition, Equation (121) can be shown as the following case analysis: if $\gamma\le\frac{L_{xy}^2}{\mu_{xy}^2}+1$ then

$$\left|1-\frac{2L_{xy}^2}{\gamma\mu_{xy}^2}\right|^2-\left(1-\frac{4(\gamma-1)}{\gamma^2}\right)=\frac{4}{\gamma^2}\left(\frac{L_{xy}^2}{\mu_{xy}^2}-1\right)\left(\frac{L_{xy}^2}{\mu_{xy}^2}+1-\gamma\right)\ge 0,$$

if $\frac{L_{xy}^2}{\mu_{xy}^2}+1<\gamma\le\frac{2L_{xy}^2}{\mu_{xy}^2}$ then

$$\frac{2L_{xy}}{\gamma\mu_{xy}}\sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}-1}+\left(\frac{2L_{xy}^2}{\gamma\mu_{xy}^2}-1\right)-\left(1-\frac{2}{\gamma}\right)\ge\frac{2}{\gamma}\left\{\frac{L_{xy}}{\mu_{xy}}\sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}-1}-\left(\frac{L_{xy}^2}{\mu_{xy}^2}-1\right)\right\}\ge 0,$$

and if $\gamma>\frac{2L_{xy}^2}{\mu_{xy}^2}$,

$$\frac{2L_{xy}}{\gamma\mu_{xy}}\sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}-1}+\left(1-\frac{2L_{xy}^2}{\gamma\mu_{xy}^2}\right)-\left(1-\frac{2}{\gamma}\right)=\frac{2}{\gamma}\left\{\frac{L_{xy}}{\mu_{xy}}\sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}-1}-\left(\frac{L_{xy}^2}{\mu_{xy}^2}-1\right)\right\}\ge 0.$$

Solving Equation (119),

$$1-(\gamma-1)\mu_{xy}^2X=1-\gamma L_{xy}^2X+\frac{\gamma^2L_{xy}^4}{4}X^2+\frac{\gamma^2L_{xy}^4X^2-4L_{xy}^2X}{4}+2\left|1-\frac{\gamma L_{xy}^2X}{2}\right|\frac{\sqrt{\gamma^2L_{xy}^4X^2-4L_{xy}^2X}}{2},$$

$$\left\{(\gamma+1)L_{xy}^2-(\gamma-1)\mu_{xy}^2\right\}X-\frac{\gamma^2L_{xy}^4}{2}X^2=\left|1-\frac{\gamma L_{xy}^2X}{2}\right|\sqrt{\gamma^2L_{xy}^4X^2-4L_{xy}^2X},$$

$$\begin{aligned}&\left\{(\gamma+1)L_{xy}^2-(\gamma-1)\mu_{xy}^2\right\}^2X^2-\left\{(\gamma+1)L_{xy}^2-(\gamma-1)\mu_{xy}^2\right\}\gamma^2L_{xy}^4X^3+\frac{\gamma^4L_{xy}^8}{4}X^4\\&=\left(1-\gamma L_{xy}^2X+\frac{\gamma^2L_{xy}^4X^2}{4}\right)\left(\gamma^2L_{xy}^4X^2-4L_{xy}^2X\right)\\&=-4L_{xy}^2X+\left(\gamma^2+4\gamma\right)L_{xy}^4X^2-\left(\gamma^3+\gamma^2\right)L_{xy}^6X^3+\frac{\gamma^4L_{xy}^8}{4}X^4,\end{aligned}$$

$$4L_{xy}^2-\underbrace{\left((2\gamma-1)L_{xy}^4+2(\gamma^2-1)L_{xy}^2\mu_{xy}^2-(\gamma-1)^2\mu_{xy}^4\right)}_{=:B(\gamma)}X+(\gamma^3-\gamma^2)L_{xy}^4\mu_{xy}^2X^2=0. \tag{122}$$

The discriminant equals

$$\begin{aligned}&B(\gamma)^2-16(\gamma^3-\gamma^2)L_{xy}^6\mu_{xy}^2\\&=\left(L_{xy}^2-(\gamma-1)\mu_{xy}^2\right)^2\left((2\gamma-1)^2L_{xy}^4-2L_{xy}^2\mu_{xy}^2(2\gamma^2-\gamma-1)+(\gamma-1)^2\mu_{xy}^4\right)\\&=\left(L_{xy}^2-(\gamma-1)\mu_{xy}^2\right)^2\left\{\left((2\gamma+1)L_{xy}^2-(\gamma-1)\mu_{xy}^2\right)^2-(2\sqrt{2\gamma}L_{xy}^2)^2\right\}\geq 0,\end{aligned} \tag{123}$$

where the last inequality is due to

$$(2\gamma+1)L_{xy}^2-(\gamma-1)\mu_{xy}^2\geq(\gamma+2)L_{xy}^2\geq 2\sqrt{2\gamma}L_{xy}^2.$$

Solving the quadratic equation in Equation (122), there are two possible optimal choices of X.

$$X=\frac{B(\gamma)\pm\sqrt{B(\gamma)^2-16(\gamma^3-\gamma^2)L_{xy}^6\mu_{xy}^2}}{2(\gamma^3-\gamma^2)L_{xy}^4\mu_{xy}^2}$$

Nevertheless, we take only the minus sign to maximize the value of $\sqrt{1-(\gamma-1)\mu_{xy}^2X}$ among possible X's. This is because, if we took the plus sign, the X would be a solution of

$$\sqrt{1-(\gamma-1)\mu_{xy}^2X}=-\left|1-\frac{\gamma L_{xy}^2X}{2}\right|+\frac{\sqrt{\gamma^2L_{xy}^4X^2-4L_{xy}^2X}}{2},$$

but would not be a solution of Equation (119). In other words, the optimal choice of X given a fixed γ is

$$X_*(\gamma):=\frac{B(\gamma)-\sqrt{B(\gamma)^2-16(\gamma^3-\gamma^2)L_{xy}^6\mu_{xy}^2}}{2(\gamma^3-\gamma^2)L_{xy}^4\mu_{xy}^2}. \tag{124}$$

Putting this into the left hand side of Equation (119), now we need to minimize

$$\sqrt{1-\frac{B(\gamma)-\sqrt{B(\gamma)^2-16(\gamma^3-\gamma^2)L_{xy}^6\mu_{xy}^2}}{2\gamma^2L_{xy}^4}}.$$

Here we utilize the following fact.

Proposition E.4. *Recall that* $B(\gamma) = (2\gamma-1)L_{xy}^4 + 2(\gamma^2-1)L_{xy}^2\mu_{xy}^2 - (\gamma-1)^2\mu_{xy}^4$. *Then,*

$$h(\gamma) := \frac{B(\gamma) - \sqrt{B(\gamma)^2 - 16(\gamma^3-\gamma^2)L_{xy}^6\mu_{xy}^2}}{\gamma^2}$$

is increasing for $\gamma \in \left[1, 1+\frac{L_{xy}^2}{\mu_{xy}^2}\right]$ *and decreasing for* $\gamma \in \left[1+\frac{L_{xy}^2}{\mu_{xy}^2}, \infty\right)$.

The proof can be found in Appendix E.3.3. Thus, the optimal value of γ is

$$\gamma_* = 1 + \frac{L_{xy}^2}{\mu_{xy}^2}.$$

In this case,

$$\begin{aligned} B(\gamma_*) &= \left(\frac{2L_{xy}^2}{\mu_{xy}^2}+1\right)L_{xy}^4 + 2\left(\frac{L_{xy}^2}{\mu_{xy}^2}+2\right)L_{xy}^4 - L_{xy}^4 \\ &= 4\left(\frac{L_{xy}^2}{\mu_{xy}^2}+1\right)L_{xy}^4, \end{aligned}$$

from which we can check

$$B(\gamma_*)^2 - 16(\gamma_*^3-\gamma_*^2)L_{xy}^6\mu_{xy}^2 = 0.$$

Therefore, the optimal X in Equation (124) becomes much simpler:

$$\begin{aligned} X_*(\gamma_*) &= \frac{B(\gamma_*)}{2(\gamma_*^3-\gamma_*^2)L_{xy}^4\mu_{xy}^2} \\ &= \frac{4\left(\frac{L_{xy}^2}{\mu_{xy}^2}+1\right)L_{xy}^4}{2\left(1+\frac{L_{xy}^2}{\mu_{xy}^2}\right)^2 L_{xy}^6} \\ &= \frac{2\mu_{xy}^2}{L_{xy}^2(L_{xy}^2+\mu_{xy}^2)}, \end{aligned}$$

and the corresponding spectral radius is

$$\begin{aligned} \sqrt{1-(\gamma_*-1)\cdot X_*(\gamma_*)\cdot\mu_{xy}^2} &= \sqrt{1-\frac{2\mu_{xy}^2}{L_{xy}^2+\mu_{xy}^2}} \\ &= \sqrt{\frac{L_{xy}^2-\mu_{xy}^2}{L_{xy}^2+\mu_{xy}^2}} \end{aligned}$$

which is an even better (*i.e.*, smaller) spectral radius than those in **Case 1** and **Case 2**. This concludes the proof. □

E.3. Proofs used in Appendix E

Here we prove some technical propositions and lemmas used throughout Appendix E.

E.3.1. Proof of Proposition E.1

Here we prove Proposition E.1, restated below for the sake of readability.

Proposition E.1. *The polynomial* $P_i(\lambda)$ *defined in Equation* (113) *has no positive real root if*

$$\psi_i|\Gamma-\Delta| \le \min\left\{(1-\Gamma)^2, (1-\Delta)^2\right\}.$$

Proof. Without loss of generality, suppose $0 \le \Gamma \le \Delta < 1$. Let $p(\lambda) = \lambda(\lambda-1)^2$ and $q(\lambda) = -\psi_i(\lambda-\Gamma)(\lambda-\Delta)$; simply $P_i(\lambda) = p(\lambda) - q(\lambda)$. Since $p(\lambda) \ge 0$ for $\lambda \ge 0$ and $q(\lambda) \ge 0$ only if $\lambda \in [\Gamma, \Delta] \subset [0,1]$, $P_i(\lambda)$ can have a positive root only in the interval $[\Gamma, \Delta]$. So it suffices to show that $P_i(\lambda) > 0$ for $\lambda \in [\Gamma, \Delta]$ for proving the proposition.

Note that, for $\lambda \in [\Gamma, \Delta]$, $P_i(\lambda) \ge \lambda(1-\Delta)^2 + \psi_i(\lambda-\Gamma)(\lambda-\Delta) =: Q(\lambda)$. Now it suffices to show $Q(\lambda) > 0$ for $\lambda \in [\Gamma, \Delta]$.

Note that $Q(\lambda)$ is a quadratic polynomial and

$$Q(\lambda) = \psi_i \left(\lambda - \frac{\psi_i(\Gamma+\Delta) - (1-\Delta)^2}{2\psi_i} \right)^2 - \frac{\left\{\psi_i(\Gamma+\Delta) - (1-\Delta)^2\right\}^2}{4\psi_i} + \psi_i \Gamma\Delta.$$

Since $0 < Q(\Gamma) = \Gamma(1-\Delta)^2 \le Q(\Delta) = \Delta(1-\Delta)^2$, we can ensure $Q(\lambda) > 0$ on $[\Gamma, \Delta]$ if $\frac{\psi_i(\Gamma+\Delta)-(1-\Delta)^2}{2\psi_i} \le \Gamma$. It is equivalent to $\psi_i(\Delta - \Gamma) \le (1-\Delta)^2$, which proves the proposition. □

E.3.2. Proof of Proposition E.2

Here we prove Proposition E.2, restated below for the sake of readability.

Proposition E.2.

$$\Gamma + \Delta - \frac{\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)} \left(2 - \psi_i + \frac{\psi_i \Gamma\Delta}{1-\psi_i(\Gamma+\Delta)} \right) \ge \frac{1}{4}(\Gamma + \Delta - 2\Gamma\Delta) > 0$$

if $\psi_i \le \frac{\Gamma+\Delta-2\Gamma\Delta}{2(\Gamma+\Delta)^2}$.

Proof. Since $1 - \psi_i(\Gamma+\Delta) \in (0, 1]$, the left hand side can be lower bounded as

$$\begin{aligned}
&\Gamma + \Delta - \frac{\Gamma\Delta}{1-\psi_i(\Gamma+\Delta)} \left(2 - \psi_i + \frac{\psi_i \Gamma\Delta}{1-\psi_i(\Gamma+\Delta)} \right) \\
&= \frac{(\Gamma+\Delta)(1-\psi_i(\Gamma+\Delta))^2 - \Gamma\Delta\left((2-\psi_i)(1-\psi_i(\Gamma+\Delta)) + \psi_i\Gamma\Delta\right)}{(1-\psi_i(\Gamma+\Delta))^2} \\
&\ge (\Gamma+\Delta)(1-\psi_i(\Gamma+\Delta))^2 - \Gamma\Delta\left((2-\psi_i)(1-\psi_i(\Gamma+\Delta)) + \psi_i\Gamma\Delta\right),
\end{aligned} \tag{125}$$

Which is a quadratic polynomial of ψ_i. Let

$$\begin{aligned}
R(x) &:= (\Gamma+\Delta)(1-x(\Gamma+\Delta))^2 - \Gamma\Delta\left((2-x)(1-x(\Gamma+\Delta)) + x\Gamma\Delta\right) \\
&= \underbrace{\left\{(\Gamma+\Delta)\left(\Gamma^2+\Gamma\Delta+\Delta^2\right)\right\}}_{=a>0} x^2 - \underbrace{\left\{(\Gamma+\Delta-\Gamma\Delta)^2 + \Gamma^2 + \Gamma\Delta + \Delta^2\right\}}_{=b>0} x + \underbrace{\left\{\Gamma+\Delta-2\Gamma\Delta\right\}}_{=c>0}.
\end{aligned}$$

The discriminant of $R(x)$ is

$$\begin{aligned}
D &= b^2 - 4ac \\
&= \left\{(\Gamma+\Delta-\Gamma\Delta)^2 + \Gamma^2 + \Gamma\Delta + \Delta^2\right\}^2 - 4(\Gamma+\Delta)\left(\Gamma^2+\Gamma\Delta+\Delta^2\right)\left\{\Gamma+\Delta-2\Gamma\Delta\right\} \\
&= \Gamma^2\Delta^2\left\{8(\Gamma+\Delta)^2 + (-1+\Gamma\Delta)^2 - 4(\Gamma+\Delta)(1+\Gamma\Delta)\right\} \\
&= \Gamma^2\Delta^2\left\{\Gamma^2(\Delta^2-4\Delta+8) - 2\Gamma(2\Delta^2-7\Delta+2) + 8\Delta^2 - 4\Delta + 1\right\} \\
&= \Gamma^2\Delta^2\left\{(\Delta^2-4\Delta+8)\left(\Gamma - \frac{2\Delta^2-7\Delta+2}{\Delta^2-4\Delta+8}\right)^2 + \frac{(\Delta^2-4\Delta+8)(8\Delta^2-4\Delta+1) - (2\Delta^2-7\Delta+2)^2}{\Delta^2-4\Delta+8}\right\} \\
&= \Gamma^2\Delta^2\left\{((\Delta-2)^2+4)\left(\Gamma - \frac{2\Delta^2-7\Delta+2}{\Delta^2-4\Delta+8}\right)^2 + \frac{4(\Delta^2+1)(\Delta-1)^2 + 16\Delta^2}{(\Delta-2)^2+4}\right\} \ge 0,
\end{aligned}$$

so $R(x)$ must have two (possibly identical) positive real roots. This means that if we find a lower bound $\bar{x} > 0$ for the roots, we can confirm that $R(x) \geq R(\bar{x})$ for all $x \in [0, \bar{x}]$. Using the fact $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for all $x \leq 1$, we have

$$\begin{aligned}
\frac{b - \sqrt{b^2 - 4ac}}{2a} &\geq \frac{b}{2a}\left(1 - 1 + \frac{2ac}{b^2}\right) \\
&= \frac{c}{b} \\
&= \frac{\Gamma + \Delta - 2\Gamma\Delta}{(\Gamma + \Delta - \Gamma\Delta)^2 + \Gamma^2 + \Gamma\Delta + \Delta^2} \\
&> \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2} =: \bar{x}.
\end{aligned}$$

Continuing from Equation (125), since we assumed $\psi_i \leq \bar{x}$,

$$\begin{aligned}
&R(\psi_i) \geq R(\bar{x}) \\
&= (\Gamma + \Delta)\left(1 - \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)}\right)^2 - \Gamma\Delta\left(\left(2 - \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2}\right)\left(1 - \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)}\right) + \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2}\Gamma\Delta\right) \\
&= \frac{(\Gamma + \Delta + 2\Gamma\Delta)^2}{4(\Gamma + \Delta)} - \Gamma\Delta\left(\left(2 - \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2}\right)\frac{\Gamma + \Delta + 2\Gamma\Delta}{2(\Gamma + \Delta)} + \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2}\Gamma\Delta\right) \\
&\geq \frac{(\Gamma + \Delta + 2\Gamma\Delta)^2}{4(\Gamma + \Delta)} - \Gamma\Delta\left(\frac{\Gamma + \Delta + 2\Gamma\Delta}{\Gamma + \Delta} + \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2}\Gamma\Delta\right) \\
&= \frac{(\Gamma + \Delta + 2\Gamma\Delta)(\Gamma + \Delta - 2\Gamma\Delta)}{4(\Gamma + \Delta)} - \frac{\Gamma + \Delta - 2\Gamma\Delta}{2(\Gamma + \Delta)^2}\Gamma^2\Delta^2 \\
&= \frac{\Gamma + \Delta - 2\Gamma\Delta}{4(\Gamma + \Delta)^2}\left\{(\Gamma + \Delta + 2\Gamma\Delta)(\Gamma + \Delta) - 2\Gamma^2\Delta^2\right\} \\
&\geq \frac{1}{4}(\Gamma + \Delta - 2\Gamma\Delta) > 0.
\end{aligned}$$

which concludes the proof of the proposition. □

E.3.3. Proof of Proposition E.4

Here we prove Proposition E.4, restated below for the sake of readability.

Proposition E.4. *Recall that* $B(\gamma) = (2\gamma - 1)L_{xy}^4 + 2(\gamma^2 - 1)L_{xy}^2\mu_{xy}^2 - (\gamma - 1)^2\mu_{xy}^4$. *Then,*

$$h(\gamma) := \frac{B(\gamma) - \sqrt{B(\gamma)^2 - 16(\gamma^3 - \gamma^2)L_{xy}^6\mu_{xy}^2}}{\gamma^2}$$

is increasing for $\gamma \in \left[1, 1 + \frac{L_{xy}^2}{\mu_{xy}^2}\right]$ *and decreasing for* $\gamma \in \left[1 + \frac{L_{xy}^2}{\mu_{xy}^2}, \infty\right)$.

Proof. From the calculation in Equation (123),

$$\begin{aligned}
h(\gamma) &= \frac{B(\gamma) - \left|L_{xy}^2 - (\gamma - 1)\mu_{xy}^2\right|\sqrt{(2\gamma - 1)^2L_{xy}^4 - 2L_{xy}^2\mu_{xy}^2(2\gamma^2 - \gamma - 1) + (\gamma - 1)^2\mu_{xy}^4}}{\gamma^2} \\
&= \min\left\{F(\gamma), G(\gamma)\right\}
\end{aligned}$$

where

$$\begin{aligned}
F(\gamma) &= \frac{B(\gamma) - \left(L_{xy}^2 - (\gamma - 1)\mu_{xy}^2\right)\sqrt{(2\gamma - 1)^2L_{xy}^4 - 2L_{xy}^2\mu_{xy}^2(2\gamma^2 - \gamma - 1) + (\gamma - 1)^2\mu_{xy}^4}}{\gamma^2}, \\
G(\gamma) &= \frac{B(\gamma) + \left(L_{xy}^2 - (\gamma - 1)\mu_{xy}^2\right)\sqrt{(2\gamma - 1)^2L_{xy}^4 - 2L_{xy}^2\mu_{xy}^2(2\gamma^2 - \gamma - 1) + (\gamma - 1)^2\mu_{xy}^4}}{\gamma^2}.
\end{aligned}$$

We want to show that $F(\gamma)$ is increasing and $G(\gamma)$ is decreasing for $\gamma \in [1, \infty)$. Let

$$\begin{aligned}
J(\gamma) &:= \frac{B(\gamma)}{\gamma^2} = \left(\frac{2\gamma-1}{\gamma^2}\right) L_{xy}^4 + 2\left(1-\frac{1}{\gamma^2}\right) L_{xy}^2\mu_{xy}^2 - \left(1-\frac{1}{\gamma}\right)^2 \mu_{xy}^4, \\
K(\gamma) &:= \frac{L_{xy}^2 - (\gamma-1)\mu_{xy}^2}{\gamma^2}, \\
M(\gamma) &:= (2\gamma-1)^2 L_{xy}^4 - 2L_{xy}^2\mu_{xy}^2(2\gamma^2-\gamma-1) + (\gamma-1)^2\mu_{xy}^4,
\end{aligned}$$

so that

$$\begin{aligned}
F(\gamma) &= J(\gamma) - K(\gamma)\sqrt{M(\gamma)}, \\
G(\gamma) &= J(\gamma) + K(\gamma)\sqrt{M(\gamma)}.
\end{aligned}$$

Then,

$$\begin{aligned}
J'(\gamma) &= -2\left(\frac{\gamma-1}{\gamma^3}\right)(L_{xy}^4 + \mu_{xy}^4) + \frac{4}{\gamma^3} L_{xy}^2\mu_{xy}^2, \\
K'(\gamma) &= \frac{-2L_{xy}^2 + (\gamma-2)\mu_{xy}^2}{\gamma^3}, \\
M'(\gamma) &= 4(2\gamma-1)L_{xy}^4 - 2(4\gamma-1)L_{xy}^2\mu_{xy}^2 + 2(\gamma-1)\mu_{xy}^4.
\end{aligned}$$

So,

$$\begin{aligned}
F'(\gamma) &= \frac{2J'(\gamma)\sqrt{M(\gamma)} - 2K'(\gamma)M(\gamma) - K(\gamma)M'(\gamma)}{2\sqrt{M(\gamma)}}, \\
G'(\gamma) &= \frac{2J'(\gamma)\sqrt{M(\gamma)} + 2K'(\gamma)M(\gamma) + K(\gamma)M'(\gamma)}{2\sqrt{M(\gamma)}}.
\end{aligned}$$

We proceed the calculation with $\kappa := \frac{L_{xy}}{\mu_{xy}} \geq 1$.

$$\begin{aligned}
\frac{\gamma^3}{2\mu_{xy}^6} F'(\gamma)\sqrt{M(\gamma)} &= \frac{\gamma^3}{\mu_{xy}^6}\left(\frac{1}{2}J'(\gamma)\sqrt{M(\gamma)} - \frac{1}{2}K'(\gamma)M(\gamma) - \frac{1}{4}K(\gamma)M'(\gamma)\right) \\
&= \left(-(\gamma-1)(\kappa^4+1) + 2\kappa^2\right)\sqrt{(2\gamma-1)^2\kappa^4 - 2(2\gamma^2-\gamma-1)\kappa^2 + (\gamma-1)^2} \\
&\quad - \frac{1}{2}\left(-2\kappa^2 + (\gamma-2)\right)\left((2\gamma-1)^2\kappa^4 - 2(2\gamma^2-\gamma-1)\kappa^2 + (\gamma-1)^2\right) \\
&\quad - \frac{\gamma}{2}\left(\kappa^2 - (\gamma-1)\right)\left((4\gamma-2)\kappa^4 - (4\gamma-1)\kappa^2 + (\gamma-1)\right) \\
&= \left(-(\kappa^4+1)\gamma + (\kappa^2+1)^2\right)\sqrt{(2\kappa^2-1)^2\gamma^2 - (4\kappa^4-2\kappa^2+2)\gamma + (\kappa^2+1)^2} \\
&\quad + (2\kappa^6 + \kappa^4 - 2\kappa^2 + 1)\gamma^2 - (\kappa^2+1)(3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^3.
\end{aligned} \tag{126}$$

We show that this is indeed nonnegative for $\gamma \geq 1$ and $\kappa \geq 1$. To this end, note that,

$$\begin{aligned}
&(2\kappa^6 + \kappa^4 - 2\kappa^2 + 1)\gamma^2 - (\kappa^2+1)(3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^3 \\
&= 2\gamma^2 + (\kappa^2+1)\left\{(2\kappa^4 - \kappa^2 - 1)\gamma^2 - (3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^2\right\} \\
&\geq 2\left\{(2\kappa^4 - \kappa^2)\gamma^2 - (3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^2\right\} \\
&= 2\left\{(2\kappa^4 - \kappa^2)\left(\gamma - \frac{3\kappa^4 - \kappa^2 + 2}{4\kappa^4 - 2\kappa^2}\right)^2 - \frac{(3\kappa^4 - \kappa^2 + 2)^2 - 4(\kappa^2+1)^2(2\kappa^4 - \kappa^2)}{4(2\kappa^4 - \kappa^2)}\right\} \\
&= 2\left\{(2\kappa^4 - \kappa^2)\left(\gamma - \frac{3\kappa^4 - \kappa^2 + 2}{4\kappa^4 - 2\kappa^2}\right)^2 - \frac{\kappa^8 - 18\kappa^6 + 13\kappa^4 + 4}{4(2\kappa^4 - \kappa^2)}\right\}.
\end{aligned}$$

Note that $2\kappa^4 - \kappa^2 > 0$. Also, (i) if $1 \le \kappa < 4$ then

$$\begin{aligned}\kappa^8 - 18\kappa^6 + 13\kappa^4 + 4 &= (\kappa-1)(\kappa+1)(\kappa^3 - 5\kappa^2 + 4\kappa - 2)(\kappa^3 + 5\kappa^2 + 4\kappa + 2) \\ &= (\kappa-1)(\kappa+1)\left((\kappa-4)(\kappa-1)\kappa - 2\right)(\kappa^3 + 5\kappa^2 + 4\kappa + 2) < 0;\end{aligned}$$

(ii) if $\kappa \ge 4$ then $\frac{3\kappa^4 - \kappa^2 + 2}{4\kappa^4 - 2\kappa^2} < 1$ and

$$(2\kappa^4 - \kappa^2)\cdot 1^2 - (3\kappa^4 - \kappa^2 + 2)\cdot 1 + (\kappa^2+1)^2 = 2(\kappa^2 - 1) + 1 > 0.$$

By (i) and (ii),

$$\begin{aligned}&(2\kappa^6 + \kappa^4 - 2\kappa^2 + 1)\gamma^2 - (\kappa^2+1)(3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^3 \\ &\ge 2\left\{(2\kappa^4 - \kappa^2)\left(\gamma - \frac{3\kappa^4 - \kappa^2 + 2}{4\kappa^4 - 2\kappa^2}\right)^2 - \frac{\kappa^8 - 18\kappa^6 + 13\kappa^4 + 4}{4(2\kappa^4 - \kappa^2)}\right\} > 0.\end{aligned} \tag{127}$$

So if $1 \le \gamma < \frac{(\kappa^2+1)^2}{\kappa^4+1}$, $\left(-(\kappa^4+1)\gamma + (\kappa^2+1)^2\right) \ge 0$, which proves the non-negativity of Equation (126). In addition, observe that the following inequality holds for all $\gamma \ge 1$:

$$\begin{aligned}&\left\{(2\kappa^6 + \kappa^4 - 2\kappa^2 + 1)\gamma^2 - (\kappa^2+1)(3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^3\right\}^2 \\ &\quad - \left((\kappa^4+1)\gamma - (\kappa^2+1)^2\right)^2\left\{(2\kappa^2-1)^2\gamma^2 - (4\kappa^4 - 2\kappa^2 + 2)\gamma + (\kappa^2+1)^2\right\} \\ &= 8\kappa^6(\kappa^2-1)^2(\gamma^4 - \gamma^3) \ge 0.\end{aligned} \tag{128}$$

This also proves the non-negativity of Equation (126) in the case of $\gamma \ge \frac{(\kappa^2+1)^2}{\kappa^4+1}$. As a result, we just showed that $F'(\gamma) \ge 0$ for $\gamma \ge 1$ and $\kappa \ge 1$. We now turn to prove $G'(\gamma) \le 0$.

$$\begin{aligned}\frac{\gamma^3}{2\mu_{xy}^6}G'(\gamma)\sqrt{M(\gamma)} &= \frac{\gamma^3}{\mu_{xy}^6}\left(\frac{1}{2}J'(\gamma)\sqrt{M(\gamma)} + \frac{1}{2}K'(\gamma)M(\gamma) + \frac{1}{4}K(\gamma)M'(\gamma)\right) \\ &= \left(-(\gamma-1)(\kappa^4+1) + 2\kappa^2\right)\sqrt{(2\gamma-1)^2\kappa^4 - 2(2\gamma^2 - \gamma - 1)\kappa^2 + (\gamma-1)^2} \\ &\quad + \frac{1}{2}\left(-2\kappa^2 + (\gamma-2)\right)\left((2\gamma-1)^2\kappa^4 - 2(2\gamma^2 - \gamma - 1)\kappa^2 + (\gamma-1)^2\right) \\ &\quad + \frac{\gamma}{2}\left(\kappa^2 - (\gamma-1)\right)\left((4\gamma-2)\kappa^4 - (4\gamma-1)\kappa^2 + (\gamma-1)\right) \\ &= \left(-(\kappa^4+1)\gamma + (\kappa^2+1)^2\right)\sqrt{(2\kappa^2-1)^2\gamma^2 - (4\kappa^4 - 2\kappa^2 + 2)\gamma + (\kappa^2+1)^2} \\ &\quad - \left\{(2\kappa^6 + \kappa^4 - 2\kappa^2 + 1)\gamma^2 - (\kappa^2+1)(3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^3\right\}.\end{aligned}$$

We show that this is nonpositive for $\gamma \ge 1$ and $\kappa \ge 1$. To this end, note that again from Equation (127),

$$(2\kappa^6 + \kappa^4 - 2\kappa^2 + 1)\gamma^2 - (\kappa^2+1)(3\kappa^4 - \kappa^2 + 2)\gamma + (\kappa^2+1)^3 \ge 0.$$

Also, if $1 \le \gamma < \frac{(\kappa^2+1)^2}{\kappa^4+1}$, Equation (128) still holds. On the other hand, if $\gamma \ge \frac{(\kappa^2+1)^2}{\kappa^4+1}$, $\left(-(\kappa^4+1)\gamma + (\kappa^2+1)^2\right) \le 0$. These indeed prove that $G'(\gamma) \le 0$ for $\gamma \ge 1$ and $\kappa \ge 1$.

Now we conclude the proof by remarking that $h(\gamma) = F(\gamma)$ if $\gamma \in \left[1, 1 + \frac{L_{xy}^2}{\mu_{xy}^2}\right]$ and $h(\gamma) = G(\gamma)$ if $\gamma \in \left[1 + \frac{L_{xy}^2}{\mu_{xy}^2}, \infty\right)$.

□

F. Proof of Proposition A.1

Here we prove Proposition A.1 of Appendix A, restated below for the sake of readability.

Proposition A.1. *There exists a 6-dimensional function $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ with $d_x = d_y = 3$ such that for any constant step sizes $\alpha, \beta > 0$, the convergence of **OGD** requires an iteration complexity of rate at least*

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log \frac{1}{\epsilon}\right)$$

in order to have $\|\boldsymbol{z}_K - \boldsymbol{z}_\star\|^2 \le \epsilon$.

Proof. Recall that **OGD** takes updates of the form:

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - 2\alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) + \alpha\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}),\\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + 2\beta\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) - \beta\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}).\end{aligned}$$

We use the same worst-case function as in Theorem 3.3:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\begin{bmatrix}x\\ s\\ t\\ y\\ u\\ v\end{bmatrix}^\top \begin{bmatrix}\mu_x & 0 & 0 & L_{xy} & 0 & 0\\ 0 & \mu_x & 0 & 0 & 0 & 0\\ 0 & 0 & L_x & 0 & 0 & 0\\ L_{xy} & 0 & 0 & -\mu_y & 0 & 0\\ 0 & 0 & 0 & 0 & -\mu_y & 0\\ 0 & 0 & 0 & 0 & 0 & -L_y\end{bmatrix}\begin{bmatrix}x\\ s\\ t\\ y\\ u\\ v\end{bmatrix},$$

where $\boldsymbol{x} = (x, s, t)$ and $\boldsymbol{y} = (y, u, v)$. It can be easily checked that f is a quadratic function (*i.e.,* Hessian is constant) such that $f \in \mathcal{F}(\mu_x, \mu_y, L_x, L_y, L_{xy})$ and $\boldsymbol{x}_\star = \boldsymbol{y}_\star = \boldsymbol{0} \in \mathbb{R}^3$.

Let us define

$$\boldsymbol{A} = \begin{bmatrix}\mu_x & 0 & 0\\ 0 & \mu_x & 0\\ 0 & 0 & L_x\end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix}L_{xy} & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix}\mu_y & 0 & 0\\ 0 & \mu_y & 0\\ 0 & 0 & L_y\end{bmatrix}.$$

We first observe that the k-th step of **OGD** satisfies

$$\begin{bmatrix}\boldsymbol{x}_{k+1}\\ \boldsymbol{y}_{k+1}\\ \boldsymbol{x}_k\\ \boldsymbol{y}_k\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - 2\alpha\boldsymbol{A} & -2\alpha\boldsymbol{B} & \alpha\boldsymbol{A} & \alpha\boldsymbol{B}\\ 2\beta\boldsymbol{B}^\top & \boldsymbol{I} - 2\beta\boldsymbol{C} & -\beta\boldsymbol{B}^\top & \beta\boldsymbol{C}\\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_k\\ \boldsymbol{y}_k\\ \boldsymbol{x}_{k-1}\\ \boldsymbol{y}_{k-1}\end{bmatrix}.$$

Then the coordinate-wise updates on the k-th step of **OGD** must be

$$\begin{bmatrix}x_{k+1}\\ y_{k+1}\\ x_k\\ y_k\end{bmatrix} = \underbrace{\begin{bmatrix}1 - 2\alpha\mu_x & -2\alpha L_{xy} & \alpha\mu_x & \alpha L_{xy}\\ 2\beta L_{xy} & 1 - 2\beta\mu_y & -\beta L_{xy} & \beta\mu_y\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}}_{\triangleq \boldsymbol{P}}\begin{bmatrix}x_k\\ y_k\\ x_{k-1}\\ y_{k-1}\end{bmatrix}, \tag{129}$$

$$s_{k+1} = (1 - 2\alpha\mu_x)s_k + \alpha\mu_x s_{k-1}, \tag{130}$$

$$t_{k+1} = (1 - 2\alpha L_x)t_k + \alpha L_x t_{k-1}, \tag{131}$$

$$u_{k+1} = (1 - 2\beta\mu_y)u_k + \beta\mu_y u_{k-1}, \tag{132}$$

$$v_{k+1} = (1 - 2\beta L_y)v_k + \beta L_y v_{k-1}. \tag{133}$$

First, observing that the quadratic $w^2-(1-2c)w-c=0$ has (real) roots given by

$$w=\frac{(1-2c)\pm\sqrt{(1-2c)^2+4c}}{2},$$

a recurrence relation of the form $w_{k+1}=(1-2c)w_k+cw_{k-1}$ converges if and only if

$$r=\frac{|1-2c|+\sqrt{(1-2c)^2+4c}}{2}<1,$$

which is again equivalent to $0<c<\frac{2}{3}$.

Moreover, if $0<c\leq\frac{1}{2}$, then we have

$$\begin{aligned}\frac{1}{1-r}&=\frac{1}{1-\frac{1-2c+\sqrt{(1-2c)^2+4c}}{2}}\\&=\frac{2}{1+2c-\sqrt{1+4c^2}}=\frac{1+2c+\sqrt{1+4c^2}}{2c}\geq\frac{1}{2c}=\Omega\left(\frac{1}{c}\right),\end{aligned}$$

while if $\frac{1}{2}<c<\frac{2}{3}$, then we have

$$\begin{aligned}\frac{1}{1-r}&=\frac{1}{1-\frac{2c-1+\sqrt{(1-2c)^2+4c}}{2}}\\&=\frac{2}{3-2c-\sqrt{1+4c^2}}\geq 2+\sqrt{2}\geq\left(1+\frac{1}{\sqrt{2}}\right)\cdot\frac{1}{c}=\Omega\left(\frac{1}{c}\right)\end{aligned}$$

which is because $\frac{2}{3-2c-\sqrt{1+4c^2}}$ is an increasing function in $[\frac{1}{2},\frac{2}{3})$.

For the convergence of iterations (131) and (133), the step sizes α and β are required to satisfy

$$0<\alpha L_x<\frac{2}{3}\quad\text{and}\quad 0<\beta L_y<\frac{2}{3},\tag{134}$$

by setting $c=\alpha L_x$ and/or $c=\beta L_y$.

Also, to guarantee $\|\boldsymbol{x}_K\|^2+\|\boldsymbol{y}_K\|^2<\epsilon$, we need from (130) and (132) that $s_K^2<\mathcal{O}(\epsilon)$ and $u_K^2<\mathcal{O}(\epsilon)$, respectively.

The two necessary conditions $s_K^2<\mathcal{O}(\epsilon)$ and $u_K^2<\mathcal{O}(\epsilon)$ require an iteration number of at least:

$$K=\Omega\left(\left(\frac{1}{\alpha\mu_x}+\frac{1}{\beta\mu_y}\right)\cdot\log\frac{1}{\epsilon}\right),\tag{135}$$

by setting $c=\alpha\mu_x$ and/or $c=\beta\mu_y$.

Note that (134) automatically yields

$$\frac{1}{\alpha\mu_x}+\frac{1}{\beta\mu_y}=\Omega(\kappa_x+\kappa_y).\tag{136}$$

Now, in order to ensure convergence of iteration (129), we need the matrix $\boldsymbol{P}$ to have a spectral radius smaller than one. Hence it suffices to show that $\rho(\boldsymbol{P})<1$ implies $\frac{1}{\alpha\mu_x}+\frac{1}{\beta\mu_y}=\Omega(\kappa_{xy})$.

Suppose that λ is an eigenvalue of $\boldsymbol{P}$. Then we must have

$$\det(\lambda\boldsymbol{I}-\boldsymbol{P})=\begin{vmatrix}(1-\lambda)-2\alpha\mu_x & -2\alpha L_{xy} & \alpha\mu_x & \alpha L_{xy}\\ 2\beta L_{xy} & (1-\lambda)-2\beta\mu_y & -\beta L_{xy} & \beta\mu_y\\ 1 & 0 & -\lambda & 0\\ 0 & 1 & 0 & -\lambda\end{vmatrix}=0.$$

First, we observe that $\lambda \neq 0$, since if we plug in $\lambda = 0$ we have

$$\det(\lambda \boldsymbol{I} - \boldsymbol{P}) = \det(\boldsymbol{P}) = \alpha\beta(\mu_x\mu_y + L_{xy}^2) > 0.$$

Therefore we can compute

$$\begin{aligned}
&\begin{vmatrix} (1-\lambda) - 2\alpha\mu_x & -2\alpha L_{xy} & \alpha\mu_x & \alpha L_{xy} \\ 2\beta L_{xy} & (1-\lambda) - 2\beta\mu_y & -\beta L_{xy} & \beta\mu_y \\ 1 & 0 & -\lambda & 0 \\ 0 & 1 & 0 & -\lambda \end{vmatrix} \\
&= \frac{1}{\lambda^2} \begin{vmatrix} \lambda(1-\lambda) - 2\lambda\alpha\mu_x & -2\lambda\alpha L_{xy} & \alpha\mu_x & \alpha L_{xy} \\ 2\lambda\beta L_{xy} & \lambda(1-\lambda) - 2\lambda\beta\mu_y & -\beta L_{xy} & \beta\mu_y \\ \lambda & 0 & -\lambda & 0 \\ 0 & \lambda & 0 & -\lambda \end{vmatrix} \\
&= \frac{1}{\lambda^2} \begin{vmatrix} \lambda(1-\lambda) - (2\lambda-1)\alpha\mu_x & -(2\lambda-1)\alpha L_{xy} & \alpha\mu_x & \alpha L_{xy} \\ (2\lambda-1)\beta L_{xy} & \lambda(1-\lambda) - (2\lambda-1)\beta\mu_y & -\beta L_{xy} & \beta\mu_y \\ 0 & 0 & -\lambda & 0 \\ 0 & 0 & 0 & -\lambda \end{vmatrix} \\
&= \begin{vmatrix} \lambda(1-\lambda) - (2\lambda-1)\alpha\mu_x & -(2\lambda-1)\alpha L_{xy} \\ (2\lambda-1)\beta L_{xy} & \lambda(1-\lambda) - (2\lambda-1)\beta\mu_y \end{vmatrix} \\
&= \left(\lambda(1-\lambda) - (2\lambda-1)\alpha\mu_x\right)\left(\lambda(1-\lambda) - (2\lambda-1)\beta\mu_y\right) + (2\lambda-1)^2\alpha\beta L_{xy}^2.
\end{aligned}$$

If we substitute $a = \alpha\mu_x$ and $b = \beta\mu_y$, then $\det(\lambda \boldsymbol{I} - \boldsymbol{P}) = 0$ is equivalent to

$$\left(-\lambda^2 + (1-2a)\lambda + a\right)\left(-\lambda^2 + (1-2b)\lambda + b\right) + (2\lambda-1)^2 ab\kappa_{xy}^2 = 0, \tag{137}$$

where we note that $\alpha\beta L_{xy}^2 = ab\kappa_{xy}^2$.

Hence we have a quartic equation of the form $\lambda^4 - p\lambda^3 + q\lambda^2 - r\lambda + \ell$ with coefficients given by

$$\begin{aligned}
p &= 2 - 2(a+b), \\
q &= 1 - 3a - 3b + 4ab(\kappa_{xy}^2 + 1), \\
r &= -a - b + 4ab(\kappa_{xy}^2 + 1), \\
\ell &= ab(\kappa_{xy}^2 + 1).
\end{aligned} \tag{138}$$

Note that we obviously have $p, q, r, \ell > 0$.

There exists a well-known characterization of quartic polynomials having roots with absolute values less than one.

Proposition F.1 (Grove & Ladas (2004), Theorem 1.5)**.** *Consider a quartic polynomial $x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, where a_0, a_1, a_2, a_3 are real numbers. Then a necessary and sufficient condition that all roots of the polynomial are contained in the open disk $|x| < 1$ is*

$$\begin{aligned}
&|a_1 + a_3| < 1 + a_0 + a_2, \quad |a_1 - a_3| < 2(1 - a_0), \quad a_2 - 3a_0 < 3, \\
&a_0 + a_2 + a_0^2 + a_1^2 + a_0^2a_2 + a_0a_3^2 < 1 + 2a_0a_2 + a_1a_3 + a_0a_1a_3 + a_0^3.
\end{aligned} \tag{139}$$

Also, the following corollary suggests that the coefficients are all bounded (by constants) for such cases.

Corollary F.2. *For coefficients a_0, a_1, a_2, a_3 satisfying (139), we have $|a_3| < 6$, $|a_2| < 6$, $|a_1| < 6$, $|a_0| < 1$.*

Proof. From the first three conditions, we can observe that

$$0 < 1 + a_0 + a_2, \quad 0 < 2(1 - a_0), \quad a_2 - 3a_0 < 3.$$

Hence (a_0, a_2) must be inside a triangle with endpoints $(-1, 0)$, $(1, -2)$, $(1, 6)$, which implies $|a_2| < 6$, $|a_0| < 1$.

Using this, we can also observe from the first two conditions that

$$|a_1 + a_3| < 1 + a_0 + a_2 < 8, \quad |a_1 - a_3| < 2(1 - a_0) < 4.$$

Hence (a_1, a_3) must be inside a rectangle with endpoints $(2, 6), (6, 2), (-2, -6), (-6, -2)$, implying $|a_3| < 6$, $|a_1| < 6$. □

By Corollary F.2, we can observe that a necessary condition for $\rho(\boldsymbol{P}) < 1$ is that all coefficients in (138) are of order $\mathcal{O}(1)$. In particular, this implies $ab\kappa_{xy}^2 = \alpha\beta L_{xy}^2 = \mathcal{O}(1)$ in order to assure convergence, which concludes that

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} \geq \frac{2}{\sqrt{\alpha\beta\mu_x\mu_y}} = \frac{2\kappa_{xy}}{\sqrt{\alpha\beta L_{xy}^2}} = \Omega(\kappa_{xy}). \tag{140}$$

Combining (136) and (140), we have

$$\frac{1}{\alpha\mu_x} + \frac{1}{\beta\mu_y} = \Omega(\kappa_x + \kappa_y + \kappa_{xy})$$

and therefore from (135) we can show a lower bound of

$$\Omega\left((\kappa_x + \kappa_y + \kappa_{xy}) \cdot \log\frac{1}{\epsilon}\right).$$

□

G. Details of Experiments

G.1. SCSC Quadratic Game (1): Small-scale

We run experiments on the following SCSC quadratic problem:

$$f(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{U}^\top \begin{bmatrix} \mu_x & 0 & 0 \\ 0 & L_x & 0 \\ 0 & 0 & L_x \end{bmatrix} \boldsymbol{U}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{U}^\top \begin{bmatrix} L_{xy} & 0 & 0 \\ 0 & L_{xy} & 0 \\ 0 & 0 & \mu_{xy} \end{bmatrix} \boldsymbol{V}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{V}^\top \begin{bmatrix} \mu_y & 0 & 0 \\ 0 & L_y & 0 \\ 0 & 0 & L_y \end{bmatrix} \boldsymbol{V}\boldsymbol{y},$$

where $\boldsymbol{U} \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{V} \in \mathbb{R}^{3\times 3}$ are random orthogonal matrices (generated with QR-decompositions of random Gaussian matrices). For a clear demonstration of optimization trajectories in Figure 1, we set $\boldsymbol{U} = \boldsymbol{V} = \boldsymbol{I}_{3\times 3}$. For the problem parameters, we use $L_x = L_y = L_{xy} = 1$ and $\mu_x = \mu_y = \mu_{xy} = 0.2$. We run each algorithm until it reaches $\|\boldsymbol{z}_k\|^2 < \epsilon = 10^{-50}$.

Implementation of EG. We use a general form of **EG** as follows:

$$\begin{aligned} \boldsymbol{x}_{k+\frac{1}{2}} &= \boldsymbol{x}_k - \alpha_0 \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \quad & \boldsymbol{y}_{k+\frac{1}{2}} &= \boldsymbol{y}_k + \beta_0 \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha_1 \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}), \quad & \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta_1 \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}), \end{aligned}$$

where the step sizes at explorations step ($k \to k+1/2$) and at update step ($k+1/2 \to k+1$) can differ.

Implementation of OGD. Also, we use a general form of **OGD** as follows:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \alpha_0 \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) + \alpha_1 \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}), \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \beta_0 \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) - \beta_1 \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1}), \end{aligned}$$

Parameter tuning. We tuned step sizes and other parameters (like γ and δ of **Alex-GDA**) by grid search. Since this is a quadratic problem (where the local convergence analysis directly applies), following the analysis by Zhang et al. (2022), we choose μ/L^2-scale step size for **Sim-GDA** and $1/L$-scale step size for the other algorithms ($L = \max\{L_x, L_y, L_{xy}\}$, $\mu = \min\{\mu_x, \mu_y, \mu_{xy}\}$). To be more specific,

- **Sim-GDA**: (step size) $= \frac{\mu}{CL^2}$, where $C \in \{0.5, 0.51, 0.52, \cdots, 2.99, 3\}$. (If we apply $\frac{1}{L}$-scale step size, it diverges.)
- **Alt-GDA**: (step size) $= \frac{1}{CL}$, where $C \in \{1, 1.01, 1.02, \cdots, 3.99, 4\}$.
- **EG**, **OGD**: $\alpha_0 = \beta_0 = \frac{1}{C_0 L}$ and $\alpha_1 = \beta_1 = \frac{1}{C_1 L}$, where $C_0, C_1 \in \{0.5, 0.51, 0.52, \cdots, 3.99, 4\}$
- **Alex-GDA**: (step size) $= \frac{1}{CL}$, where $C \in \{1, 1.1, 1.2, \cdots, 1.9, 2\}$, and $\gamma, \delta \in \{1.1, 1.2, 1.3, \cdots, 3.9, 4\}$

G.2. SCSC Quadratic Game (2): Higher Dimension, Extensive Comparisons

We generate the SCSC quadratic problems $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ as

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

where we randomly sample the matrices $\boldsymbol{A} \in \mathbb{R}^{d_x \times d_x}$, $\boldsymbol{B} \in \mathbb{R}^{d_x \times d_y}$, $\boldsymbol{C} \in \mathbb{R}^{d_y \times d_y}$, $\boldsymbol{U} \in \mathbb{R}^{d_x \times d_x}$, and $\boldsymbol{V} \in \mathbb{R}^{d_y \times d_y}$:

$$\begin{aligned} \boldsymbol{A} &= \mathrm{diag}(a_1, \ldots, a_{d_x}), \quad & a_1 &= \mu_x, \quad & a_2 &= L_x, \quad & a_i &\sim \mathrm{Uniform}(\mu_x, L_x), \quad & (i = 3, \ldots, d_x) \\ \boldsymbol{B} &= \mathrm{diag}(b_1, \ldots, b_{\min\{d_x, d_y\}}), \quad & b_1 &= \mu_{xy}, \quad & b_2 &= L_{xy}, \quad & b_i &\sim \mathrm{Uniform}(\mu_{xy}, L_{xy}), \quad & (i = 3, \ldots, \min\{d_x, d_y\}) \\ \boldsymbol{C} &= \mathrm{diag}(c_1, \ldots, c_{d_y}), \quad & c_1 &= \mu_y, \quad & c_2 &= L_y, \quad & c_i &\sim \mathrm{Uniform}(\mu_y, L_y), \quad & (i = 3, \ldots, d_y) \end{aligned}$$

while $\boldsymbol{U} \in \mathbb{R}^{d_x \times d_x}$ and $\boldsymbol{V} \in \mathbb{R}^{d_y \times d_y}$ are random orthogonal matrices.

For each combination of μ ($= \mu_x = \mu_y$), μ_{xy}, L ($= L_x = L_y$), and L_{xy}, we test 3 random initialization points $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ and 10 random instances of $f(\boldsymbol{x}, \boldsymbol{y})$.

Algorithms. We follow a standard implementation of heavy-ball momentum as PyTorch's implementation. We adopt Azizian et al. (2020) for **EG** with Momentum, Ramirez et al. (2023) for **OGD** with Momentum (so-called OmegaM), and Zhang & Yu (2020) for the alternating counterparts of **EG** and **OGD** (Alt-EG and Alt-OG, respectively).

Our implementation of **Alex-GDA** with momentum (Alex+M) is as in Algorithm 3.

Algorithm 3 **Alex-GDA** with Momentum

Input: Number of epochs K, step sizes $\alpha, \beta > 0$, hyperparameters $\gamma, \delta \geq 0$, momentum parameters $m_x, m_y \in \mathbb{R}$
Initialize: $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ and $\tilde{\boldsymbol{y}}_0 = \boldsymbol{y}_0 \in \mathbb{R}^{d_y}$
for $k = 0, \ldots, K-1$ **do**
$\boldsymbol{v}^x_{k+1} = m_x \boldsymbol{v}^x_k + \nabla_x f(\boldsymbol{x}_k, \tilde{\boldsymbol{y}}_k)$
$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \boldsymbol{v}^x_{k+1}$
$\tilde{\boldsymbol{x}}_{k+1} = \boldsymbol{x}_k - \gamma\alpha \boldsymbol{v}^x_{k+1}$
$\boldsymbol{v}^y_{k+1} = m_y \boldsymbol{v}^y_k + \nabla_y f(\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{y}_k)$
$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \beta \boldsymbol{v}^y_{k+1}$
$\tilde{\boldsymbol{y}}_{k+1} = \boldsymbol{y}_k + \delta\beta \boldsymbol{v}^y_{k+1}$
end for
Output: $(\boldsymbol{x}_K, \boldsymbol{y}_K) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

Computing gradient complexity. For most algorithms, the number of gradient computations equals the number of iterations. However, **EG** and its alternating counterpart (Alt-EG) take multiple gradient computations per iteration. For **EG** (with simultaneous updates), it takes two gradient computations per iteration. For Alt-EG, according to the implementation by Zhang & Yu (2020), it takes three gradient computations per iteration. Hence, we computed the gradient complexity by multiplying the number of iterations and the amount of gradient computation per iteration.

Parameter Tuning. Likewise in Appendix G.1, we choose $\frac{\mu}{\max\{L^2, L_{xy}^2\}}$-scale step size for **Sim-GDA** and $\frac{1}{\max\{L, L_{xy}\}}$-scale step size for the other algorithms. To be specific,

- **Sim-GDA** : (step size)$= \frac{C\mu}{\max\{L^2, L_{xy}^2\}}$ where $C \in \{0.1, 0.2, \ldots, 1.5\}$,
- The other algorithms (including **Sim-GDA** with momentum): (step size)$= \frac{C}{\max\{L, L_{xy}\}}$ where $C \in \{0.1, 0.2, \ldots, 1.5\}$.

We tune the momentum parameters $m_x, m_y \in \{-0.99, -0.95, -0.9, -0.8, -0.7, \ldots, 0.9, 0.95, 0.99\}$. Note that we allow the negative momentum as per the work by Gidel et al. (2019b). We tune γ and δ for **Alex-GDA** as $\gamma, \delta \in \{0.5, 0.6, 0.7, \ldots, 3.0\}$. For the momentum variant of **Alex-GDA** (Algorithm 3), we slightly reduced the range of search as $\gamma, \delta \in \{1.0, 1.1, 1.2, \ldots, 3.0\}$.

G.3. Generative Adversarial Networks: WGAN-GP

We name the combination of Adam (Kingma & Ba, 2015) and (the stochastic version of) **Sim-**/**Alt-**/**Alex-GDA** as Sim-/Alt-/Alex-Adam, respectively. In Listing 1, we provide a brief Python code based on PyTorch (Paszke et al., 2019) for GAN training with Alex-Adam. The full code base can be found at `github.com/HanseulJo/Alex-GDA/tree/main/gan`. In the code, we use the main models `netD` and `netG` (for which the weights correspond to $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively) and the auxiliary models `netD_` and `netG_`. The auxiliary models are for describing the 'tilde' variables $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$, i.e., the results of the inter-/extrapolation steps.

Learning Rates. For MNIST (Deng, 2012), we tuned the step sizes for Alex-Adam ($\{10^{-4}, 3 \times 10^{-4}\}$ for both generator and discriminator) and applied the best step size (3×10^{-4} for generator, 10^{-4} for discriminator) for the other algorithms.

For CIFAR-10 (Krizhevsky et al., 2009), we tuned the step sizes for algorithms ($\{10^{-4}, 3 \times 10^{-4}\}$ for both generator and discriminator). The best step sizes were (10^{-4} for both generator and discriminator) for Sim-Adam and (3×10^{-4} for both generator and discriminator) for Alt-Adam and Alex-Adam.

For LSUN-Bedroom 64×64 dataset (Yu et al., 2015), we fixed the step size as (10^{-4} for generator, 3×10^{-4} for discriminator) following Heusel et al. (2017).

Listing 1. PyTorch-based Python code for GAN Training with **Alex-GDA** + Adam optimizer (i.e., Alex-Adam)

```
from copy import deepcopy
import torch
from models import Discriminator, Generator  # Custom library for modeling

# Create a Discriminator (x) and a Generator (y)
netD = Discriminator(...)    # $x_0$
netG = Generator(...)        # $y_0$
netG_ = deepcopy(netG)       # $\tilde{y}_0$

# Define the optimizers
optimizerD = torch.optim.Adam(netD.parameters(), ...)
optimizerG = torch.optim.Adam(netG.parameters(), ...)

dataloader = ...     # set of real images
num_epochs = ...     # number of epochs
criterion  = ...     # loss function
label_real = ...     # label for real image e.g. all ones
label_fake = ...     # label for fake image e.g. all zeros
gamma = ...      # Alex-GDA parameter
delta = ...      # Alex-GDA parameter

for epoch in range(1, num_epochs+1):
  for data in dataloader:
    # Generate latent vectors
    noise = torch.randn(...)

    # Save $x_k$ to `netD_`
    netD_ = deepcopy(netD)

    # Compute Discriminator error: $f(x_k, \tilde{y}_k)$
    errD_real = criterion(netD(data), label_real)
    errD_fake = criterion(netD(netG_(noise)), label_fake)    # `netG_` == $\tilde{y}_k$
    errD = errD_real + errD_fake

    # Update Discriminator: $x_{k+1}$
    optimizerD.zero_grad()
    errD.backward()
    optimizerD.step()

    # Interpolation/Extrapolation step for x:
    # Compute $\tilde{x}_{k+1} = \gamma * x_{k+1} + (1-\gamma) * x_k$ and save to `netD_`
    for online, target in zip(netD.parameters(), netD_.parameters()):
      target.data = gamma * online.data + (1 - gamma) * target.data

    # Save $y_k$ to `netG_`
    netG_ = deepcopy(netG)

    # Compute Generator error: $f(\tilde{x}_{k+1}, y_k)$
    errG = criterion(netD_(netG(noise)), label_real)     # `netD_` == $\tilde{x}_{k+1}$

    # Update Generator: $y_{k+1}$
    optimizerG.zero_grad()
    errG.backward()
    optimizerG.step()

    # Interpolation/Extrapolation step for y:
    # Compute $\tilde{y}_{k+1} = \delta * y_{k+1} + (1-\delta) * y_k$ and save to `netG_`
    for online, target in zip(netG.parameters(), netG_.parameters()):
      target.data = delta * online.data + (1 - delta) * target.data
```

H. Guessing the Complexity Bound of Alt-GDA

We have found numerical evidence based on the performance estimation program (PEP) (Drori & Teboulle, 2014) that the upper complexity bound can be strictly smaller than $\mathcal{O}(\kappa^{1.5})$ for **Alt-GDA**, which we formally state in Conjecture 8.1.

Reproducing the work by Das Gupta et al. (2023), we devised a PEP-based tool that automatically optimizes the convergence rate of **Sim-**/**Alt-GDA** under SCSC and Lipschitz gradient assumptions. While the original PEP is a tool for finding the worst-case convergence rate of a *given* algorithm (with fixed and known parameters like step sizes) by solving a semidefinite programming problem, our tool tries to minimize this worst-case rate by finding optimal step sizes and optimal coefficients of the performance measure. Here, the performance measure is a linear combination of (1) the squared distance from the current iterate to the optimum, (2) the gradient norm at the current iterate, and (3) their interaction term (inner product between an iterate-optimum gap and a gradient norm), where the coefficients of the linear combination are part of optimization variables.

Using this tool, we can obtain an optimized convergence rate r of **Sim-**/**Alt-GDA** for each set of problem parameters $(\mu_x, \mu_y, L_x, L_y, L_{xy})$. (For convenience of exhibition, we set $\mu = \mu_x = \mu_y$ and $L = L_x = L_y = L_{xy}$ and define $\kappa = L/\mu$.) Recall from Equation (3) that the complexity can be expressed as $\frac{1}{1-r}$ except for the logarithmic factor. Hence, if we find how $\frac{1}{1-r}$ can be expressed as a function of κ, we will be able to guess the actual complexity in terms of κ. We draw log-log plots between $\frac{1}{1-r}$ and κ and observe its slope, which would be the exponent of κ in the complexity. Here we tune $\kappa \in \{10^1, 10^{1.2}, 10^{1.4}, \dots, 10^3\}$, and we compute the median slope of line segments, each of them connecting a pair of adjacent points.

As shown in Figure 2, the graphs for both algorithms appear close to a straight line. For **Sim-GDA**, we observe the optimal complexity is $\approx \kappa^{1.999}$: it is tight up to numerical error. On the other hand, for **Alt-GDA**, the observed lowest possible complexity is $\approx \kappa^{1.385}$ (if we utilize a pair of consecutive iterates $\boldsymbol{z}_k \to \boldsymbol{z}_{k+1}$): See Figure 2.

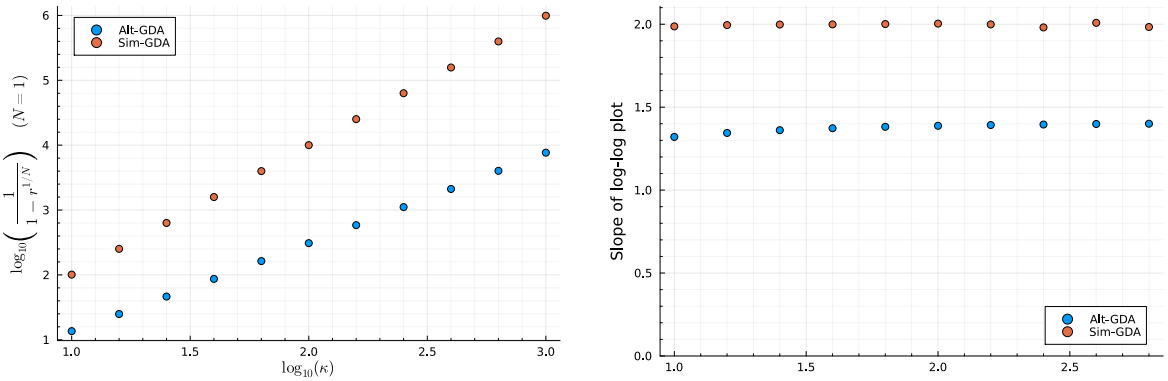

Figure 2. **Guessing the complexity bound of Sim-/Alt-GDA. Left:** log-log plot between $\kappa = L/\mu$ and the near-optimal worst-case complexity. **Right:** Slope of the log-log plot. Each point corresponds to the slope of a line segment connecting a pair of adjacent points in the left plot.

Nevertheless, we cannot assure that this *proves* that the tight complexity of **Alt-GDA** of rate $\mathcal{O}(\kappa^{1.385})$ is *tight*. This is mainly because, in fact, our tool is not perfect in terms of the function class. Although our tool implements every condition of SCSC and Lipschitz gradients as constraints of an optimization problem, it is not well understood (especially for minimax problems) whether such an implementation can properly simulate the class of SCSC functions with Lipschitz gradients; rather, it can only simulate a slightly *larger* function class including SCSC and Lipschitz-gradient functions (this is similar to the case of monotone and Lipschitz operators (Ryu et al., 2020)). Thus, the numerical value 1.385 is not tight and the true exponent can be smaller for the actual SCSC Lipschitz-gradient functions. In other words, the complexity can be smaller than $\mathcal{O}(\kappa^{1.385})$. Nonetheless, our results altogether corroborate that the upper complexity bound of **Alt-GDA** must be strictly smaller than $\mathcal{O}(\kappa^{1.5})$.