SGDA WITH SHUFFLING: FASTER CONVERGENCE FOR NONCONVEX-PŁ MINIMAX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent-ascent (SGDA) is one of the main workhorses for solving finite-sum minimax optimization problems. Most practical implementations of SGDA randomly reshuffle components and sequentially use them (*i.e.*, without-replacement sampling); however, there are few theoretical results on this approach for minimax algorithms, especially outside the easier-to-analyze (strongly-)monotone setups. To narrow this gap, we study the convergence bounds of SGDA with random reshuffling (**SGDA-RR**) for smooth nonconvex-nonconcave objectives with Polyak-Łojasiewicz (PŁ) geometry. We analyze both simultaneous and alternating SGDA-RR for *nonconvex-PŁ* and *primal-PŁ-PŁ* objectives, and obtain convergence rates faster than with-replacement SGDA. Our rates also extend to mini-batch SGDA-RR, recovering known rates for full-batch gradient descent-ascent (GDA). Lastly, we present a comprehensive lower bound for two-time-scale GDA, which matches the full-batch rate for primal-PŁ-PŁ case.

1 INTRODUCTION

A finite-sum minimax optimization problem aims to solve the following:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x};\boldsymbol{y}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x};\boldsymbol{y}), \tag{1}$$

where f_i denotes the i-th component function. In plain language, we want to minimize the average of n component functions for $\boldsymbol{x}$, while maximizing it for $\boldsymbol{y}$ given $\boldsymbol{x}$. There are many important areas in modern machine learning that fall within the minimax problem, including generative adversarial networks (GANs) (Goodfellow et al., 2020), adversarial attack and robust optimization (Madry et al., 2018; Sinha et al., 2018), multi-agent reinforcement learning (MARL) (Li et al., 2019), AUC maximization (Ying et al., 2016; Liu et al., 2020; Yuan et al., 2021), and many more. In most cases, the objective f is usually nonconvex-nonconcave, *i.e.*, neither convex in $\boldsymbol{x}$ nor concave in $\boldsymbol{y}$. Since general nonconvex-nonconcave problems are known to be intractable, we would like to tackle the problems with some additional structures, such as smoothness and Polyak-Łojasiewicz (PŁ) condition(s). We present the detailed settings for our analysis, *nonconvex-PŁ* and *primal-PŁ-PŁ* (or, *PŁ(Φ)-PŁ*), in Section 2.

One of the simplest and most popular algorithms to solve the problem (1) would be *stochastic gradient descent-ascent* (**SGDA**). This naturally extends the idea of stochastic gradient descent (SGD) used for minimization problems. Given an initial iterate $(\boldsymbol{x}_0;\boldsymbol{y}_0)$, at time $t\in\mathbb{N}$, SGDA (randomly) chooses an index $i(t)\in\{1,\dots,n\}$ and accesses the $i(t)$-th component to perform a pair of updates

$$\begin{bmatrix} \boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \alpha\nabla_1 f_{i(t)}(\boldsymbol{x}_{t-1};\boldsymbol{y}_{t-1}), \\ \boldsymbol{y}_t = \boldsymbol{y}_{t-1} + \beta\nabla_2 f_{i(t)}(\boldsymbol{x}';\boldsymbol{y}_{t-1}), \end{bmatrix} \quad \text{where } \boldsymbol{x}' = \begin{cases} \boldsymbol{x}_{t-1}, & (\textit{simSGDA}), \text{ or} \\ \boldsymbol{x}_t, & (\textit{altSGDA}). \end{cases}$$

Here, $\alpha>0$ and $\beta>0$ are the step sizes and ∇_j denotes the gradient with respect to j-th argument for $f_{i(t)}$ $(j=1,2)$. As shown in the update equations above, there are two widely used versions of SGDA: *simultaneous SGDA* (**simSGDA**), and *alternating SGDA* (**altSGDA**).

In such stochastic gradient methods, there are two main categories of sampling schemes for the component indices $i(t)$. One way is to sample $i(t)$ independently (in time) and uniformly at random from $\{1,\dots,n\}$, which is called *with-replacement sampling*. This scheme is widely adopted in

theory papers because it makes analysis of stochastic methods amenable: the noisy gradients $\nabla f_{i(t)}$ are independent over time t and are unbiased estimators of the full-batch gradient ∇f. In contrast, the vast majority of practical implementations employ *without-replacement sampling*, indicating a huge theory-practice gap. In without-replacement sampling, we sample each index precisely once at each epoch. Perhaps the most popular of such schemes is *random reshuffling* (**RR**), which uniformly randomly shuffles the order of indices at the beginning of every epoch. Unfortunately, it is well-known that without-replacement methods are much more difficult to analyze theoretically, largely because the sampled indices in each epoch are no longer independent of each other.

Interestingly, for minimization problems, several recent works overcome this obstacle and show that SGD using without-replacement sampling leads to faster convergence, given that the number of epochs is large enough (Nagaraj et al., 2019; Ahn et al., 2020; Mishchenko et al., 2020; Rajput et al., 2020; Nguyen et al., 2021; Yun et al., 2021; 2022). On the other hand, for minimax problems like (1), the majority of the studies still assume with-replacement sampling and/or rely on the independent unbiased gradient assumption (Nouiehed et al., 2019; Guo et al., 2020; Lin et al., 2020; Yan et al., 2020; Yang et al., 2020; Loizou et al., 2021; Beznosikov et al., 2022). There are very few known results on minimax optimization algorithms using without-replacement sampling, and even most of the existing ones take advantage of (strong-)convexity (in $\boldsymbol{x}$) and/or (strong-)concavity (in $\boldsymbol{y}$) of the problem (Das et al., 2022; Maheshwari et al., 2022; Yu et al., 2022). A detailed comparative analysis of these works is conducted in Section 4.

Putting all these issues into consideration, our main question is the following.

Does SGDA using without-replacement component sampling provably converge fast, even on smooth nonconvex-nonconcave objective f with PŁ structures?

1.1 Summary of our contributions

To answer the question, we analyze the convergence of SGDA with random reshuffling (**SGDA-RR**, Algorithm 1). We analyze both the simultaneous and alternating versions of SGDA-RR and prove convergence theorems for the following two regimes. Here we denote the step size ratio as $r = \beta/\alpha$.

- If $-f(\boldsymbol{x};\boldsymbol{y})$ satisfies μ_2-PŁ condition in $\boldsymbol{y}$ (**nonconvex-PŁ**) and component function f_i's are L-smooth, we prove that SGDA-RR with $r \gtrsim (L/\mu_2)^2$ converges to ε-stationarity in expectation after $\mathcal{O}\left(nrL\varepsilon^{-2} + \sqrt{n}r^{3/2}L\varepsilon^{-3}\right)$ gradient evaluations (Theorem 1).
- If we further assume μ_1-PŁ condition on the primal function $\Phi(\cdot) := \max_{\boldsymbol{y}} f(\cdot;\boldsymbol{y})$ (**primal-PŁ-PŁ**, or **PŁ(Φ)-PŁ**), we prove that SGDA-RR with $r \gtrsim (L/\mu_2)^2$ converges within ε-accuracy in expectation after $\tilde{\mathcal{O}}\left(\frac{nLr}{\mu_1}\log(\varepsilon^{-1}) + \sqrt{n}L(\frac{r}{\mu_1})^{3/2}\varepsilon^{-1}\right)$ gradient evaluations (Theorem 2).

As will be discussed in Section 4, the rates shown above are *faster* than existing results on with-replacement SGDA. In fact, Theorems 1 & 2 are special cases ($b = 1$) of our extended theorems (Theorems 4 & 5, Appendix A) that analyze *mini-batch* SGDA-RR of batch size b; by setting $b = n$, we also recover known convergence rates for full-batch gradient descent ascent (GDA). Hence, our analysis covers the entire spectrum between vanilla SGDA-RR ($b = 1$) and GDA ($b = n$).

- Additionally, we provide complexity lower bounds for solving strongly-convex-strongly-concave (SC-SC) minimax problems using two-time-scale simultaneous GDA with arbitrarily fixed step size ratio $r = \beta/\alpha$. Perhaps surprisingly, we find that the lower bound for SC-SC functions matches the convergence upper bound for a much larger class of primal-PŁ-PŁ functions when the step size ratio satisfies $r \gtrsim L^2/\mu_2^2$ (Theorem 3).

2 Problem setup

2.1 Notation

In our problem (1), the domain of every f_i is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^{d_x}$, $\mathcal{Y} = \mathbb{R}^{d_y}$, and $\mathcal{Z} = \mathbb{R}^d$: we concern unconstrained problems for simplicity. We denote the Euclidean norm and the standard inner product by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively. We often use an abbreviated notation $\boldsymbol{z} = (\boldsymbol{x};\boldsymbol{y}) \in \mathcal{Z}$ for $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$. Even when $\boldsymbol{z}$ or $(\boldsymbol{x};\boldsymbol{y})$ is followed by superscripts and/or subscripts, we use the symbols interchangeably; *e.g.*, $\boldsymbol{z}_i^k = (\boldsymbol{x}_i^k;\boldsymbol{y}_i^k)$. Note that we split the

Algorithm 1 simSGDA/altSGDA-RR

Given: The number of components n; the number of epochs K; step sizes $\alpha, \beta > 0$
Initialize: $(\boldsymbol{x}_0^1; \boldsymbol{y}_0^1) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$
for $k \in [K]$ **do**
  Sample $\sigma_k \sim \mathrm{Unif}(\mathbb{S}_n)$ ▷ RR: uniformly randomly shuffle the indices every epoch
  for $i \in [n]$ **do**
    $\boldsymbol{x}_i^k = \boldsymbol{x}_{i-1}^k - \alpha \nabla_1 f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k; \boldsymbol{y}_{i-1}^k)$
    if simSGDA-RR **then**
      $\boldsymbol{y}_i^k = \boldsymbol{y}_{i-1}^k + \beta \nabla_2 f_{\sigma_k(i)}(\boldsymbol{x}_{i-1}^k; \boldsymbol{y}_{i-1}^k)$ ▷ simultaneous update: $\boldsymbol{x}$ & $\boldsymbol{y}$
    else if altSGDA-RR **then**
      $\boldsymbol{y}_i^k = \boldsymbol{y}_{i-1}^k + \beta \nabla_2 f_{\sigma_k(i)}(\boldsymbol{x}_i^k; \boldsymbol{y}_{i-1}^k)$ ▷ alternating update: $\boldsymbol{x} \to \boldsymbol{y}$
  $(\boldsymbol{x}_0^{k+1}; \boldsymbol{y}_0^{k+1}) = (\boldsymbol{x}_n^k; \boldsymbol{y}_n^k)$

arguments $\boldsymbol{x}$ (for minimization) and $\boldsymbol{y}$ (for maximization) by a semicolon (';'). We use ∇_1 and ∇_2 to denote the gradients with respect to first and second arguments, respectively.Accordingly, we can write the full gradient as, *e.g.*, $\nabla g = [\nabla_1 g^\top; \nabla_2 g^\top]^\top$. For a positive integer N, we denote $[N] := \{1, 2, 3, \ldots, N\}$. Let the set $\mathbb{S}_N$ be a symmetric group of degree N. That is, each *permutation* $\sigma \in \mathbb{S}_N$ is a bijection from $[N]$ to itself, or equivalently, a re-arrangement of $[N]$. Lastly, we use the usual $\mathcal{O}$/Ω/Θ notation for bounds, where $\tilde{\mathcal{O}}$, $\tilde{\Omega}$, and $\tilde{\Theta}$ are used for hiding some logarithmic factors.

2.2 Algorithms: (two-time-scale) simSGDA-RR & altSGDA-RR

As we explained in Section 1, we consider simSGDA and altSGDA combined with RR, a without-replacement sampling scheme. We call them **simSGDA-RR** and **altSGDA-RR**, respectively. We present a detailed description of the methods in Algorithm 1. For completeness, we also provide an extended version that uses mini-batches of size ≥ 1 (Algorithm 2) in Appendix A. For comparison, we call the SGDA algorithms using *with-replacement* sampling by just **simSGDA** and **altSGDA**.

The quantities $\alpha, \beta > 0$ are step sizes associated with $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. We use two separate symbols α and β to allow the two step sizes to be different. Such algorithms are commonly called *two-time-scale* algorithms, in a broader sense, and they are adopted in nonconvex minimax optimization problems (Heusel et al., 2017; Lin et al., 2020; Yang et al., 2020). In fact, a recent result (Li et al., 2022) shows that having $\alpha \neq \beta$ is sometimes *necessary* for convergence.

2.3 Assumptions and definitions

To define the function classes that we are interested in solving, we introduce a few assumptions.

Assumption 1 (Component smoothness)**.** *Every i-th component $f_i : \mathcal{Z} \to \mathbb{R}$ is L-***smooth***,* i.e.*, f_i is differentiable and ∇f_i is L-Lipschitz continuous: $\|\nabla f_i(\boldsymbol{z}) - \nabla f_i(\boldsymbol{z}')\| \leq L \|\boldsymbol{z} - \boldsymbol{z}'\|$. As a result, for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$, $f_i(\boldsymbol{z}') - f_i(\boldsymbol{z}) \leq \langle \nabla f_i(\boldsymbol{z}), \boldsymbol{z}' - \boldsymbol{z} \rangle + \frac{L}{2} \|\boldsymbol{z}' - \boldsymbol{z}\|^2$.*

Since f_i's are all L-smooth, the average f of f_i's is also L-smooth.

The next one controls the discrepancy between the objective function f and its components f_i's. This is similar to Assumption 2 of Nguyen et al. (2021), adapted to minimax problems.

Assumption 2 (Component gradient variance)**.** *There exist constants $A, B \geq 0$ such that, for any $\boldsymbol{z} = (\boldsymbol{x}; \boldsymbol{y}) \in \mathcal{Z}$ and $j \in \{1, 2\}$, we have $\frac{1}{n} \sum_{i=1}^n \|\nabla_j f_i(\boldsymbol{z}) - \nabla_j f(\boldsymbol{z})\|^2 \leq A \|\nabla_j f(\boldsymbol{z})\|^2 + B$.*

In Assumption 2, letting $A = 0$ recovers a common assumption of the uniformly bounded variance of component gradients; thus, our assumption is a relaxation. Also, note that $A = B = 0$ if $n = 1$.

Assumption 3. *For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the* **primal function** *$\Phi : \mathcal{X} \to \mathbb{R}$ is well-defined as $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}; \boldsymbol{y}')$. For each $\boldsymbol{x} \in \mathcal{X}$, the set $\mathcal{Y}_{\boldsymbol{x}}^* := \arg\max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}; \boldsymbol{y}')$ is non-empty and closed. Moreover, we assume $\Phi(\boldsymbol{x})$ is bounded below by $\Phi^* = \inf_{\boldsymbol{x}' \in \mathcal{X}} \Phi(\boldsymbol{x}') > -\infty$.*

We now add an additional structure to our objective function, which is called Polyak-Łojasiewicz (PŁ) condition. A function $g : \mathbb{R}^d \to \mathbb{R}$ is said to be μ-**PŁ** if it has a minimum value g^* and satisfies

$$\|\nabla g(\boldsymbol{t})\|^2 \geq 2\mu(g(\boldsymbol{t}) - g^*). \quad (\forall\, \boldsymbol{t} \in \mathbb{R}^d)$$

Readers could find several studies and applications that the condition involved, in the papers by Karimi et al. (2016); Nouiehed et al. (2019); Yang et al. (2020); Liu et al. (2020), and more. Note

that every μ-strongly convex[1] function satisfies μ-PŁ condition, whereas a PŁ function does not need to be convex. Hence, μ-PŁ is a strict generalization of μ-strong convexity. In addition, every stationary point of a PŁ function is a global optimum, which is a benign property for optimization.

We are interested in the case where our objective function $f(\boldsymbol{x};\boldsymbol{y})$ has such a structure in terms of $\boldsymbol{y}$ (Assumption 4). Sometimes, we further assume the primal function Φ is also PŁ (Assumption 5). We emphasize that we do not necessarily assume the PŁ conditions for the individual f_i's.

Assumption 4 ($\boldsymbol{y}$-side PŁ)**.** *For each (fixed) $\boldsymbol{x} \in \mathcal{X}$, $-f(\boldsymbol{x};\cdot)$ is μ_2-PŁ,* i.e.*, for every $(\boldsymbol{x};\boldsymbol{y}) \in \mathcal{Z}$, $\|\nabla_2 f(\boldsymbol{x};\boldsymbol{y})\|^2 \geq 2\mu_2(\Phi(\boldsymbol{x}) - f(\boldsymbol{x};\boldsymbol{y}))$, where Φ is the primal function associated with f.*

Assumption 5 (Primal PŁ, or PŁ(Φ))**.** *The primal function $\Phi(\cdot) = \max_{\boldsymbol{y}'} f(\boldsymbol{x};\boldsymbol{y}')$ of f is μ_1-PŁ,* i.e.*, for every $\boldsymbol{x} \in \mathcal{X}$, $\|\nabla \Phi(\boldsymbol{x})\|^2 \geq 2\mu_1(\Phi(\boldsymbol{x}) - \Phi^*)$, where $\Phi^* = \min_{\boldsymbol{x}} \Phi(\boldsymbol{x})$ is well-defined.*

We say the function f is **nonconvex-PŁ** when it satisfies Assumption 4. Since we do not assume any convexity/concavity, it is generally hard to reach global optima. Due to the $\boldsymbol{y}$-side PŁ condition, we can guarantee that the primal function Φ is differentiable and even L_Φ-smooth with $L_\Phi \leq L + L^2/\mu_2$ (see Proposition 9 in Appendix B). Since the problem (1) can be reformulated as the minimization problem of Φ (when we can always find $\boldsymbol{y}$ well that maximizes $f(\boldsymbol{x};\boldsymbol{y})$ given $\boldsymbol{x}$), we could aim to find an approximate first-order stationary point of Φ, by making the norm of the gradient of Φ small.

On top of that, if f satisfies both Assumptions 4 and 5, the function is said to be **primal-PŁ-PŁ**, or **PŁ(Φ)-PŁ** for short.[2] In this case, we directly aim not only to decrease the primal function Φ associated with the objective function f but also to increase the function value $f(\boldsymbol{x};\boldsymbol{y})$ in terms of $\boldsymbol{y}$. To evaluate how close we are to our goal, we define a *potential function* V_λ later in Section 3. When we attain $V_\lambda(\boldsymbol{x}^*,\boldsymbol{y}^*) = 0$, it implies that we arrive at a global minimax point: $f(\boldsymbol{x}^*,\boldsymbol{y}^*) = \Phi(\boldsymbol{x}^*) = \Phi^*$. The function V_λ enables us to develop a unified analysis for nonconvex-PŁ and PŁ(Φ)-PŁ objective functions; we discuss this in greater detail in Section 3.

3 Main Results

Based on the assumptions stated in the previous section, we present the convergence results for both smooth nonconvex-PŁ objectives and smooth PŁ(Φ)-PŁ objectives. Before stating the main theorems, we first introduce the most important tool for our analyses: the *potential function*.

3.1 Potential function V_λ

For our convergence analyses, we utilize a function $V_\lambda : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined as

$$V_\lambda(\boldsymbol{x};\boldsymbol{y}) := \lambda(\Phi(\boldsymbol{x}) - \Phi^*) + (\Phi(\boldsymbol{x}) - f(\boldsymbol{x};\boldsymbol{y})), \tag{2}$$

where $\lambda > 1$ is a constant. We borrow inspiration from Yang et al. (2020) and Das et al. (2022) to come up with this function, although the placement of λ of ours is different. In fact, the convergence to a neighborhood of a global minimax point (if it exists) implies the reduction of this function. For each $\boldsymbol{x}$, a non-negative term $\Phi(\boldsymbol{x}) - f(\boldsymbol{x};\boldsymbol{y})$ gets smaller as $\boldsymbol{y}$ makes $f(\boldsymbol{x};\boldsymbol{y})$ larger. The term becomes zero when $\boldsymbol{y} = \boldsymbol{y}^*(\boldsymbol{x})$ for some $\boldsymbol{y}^*(\boldsymbol{x}) \in \mathcal{Y}^*_{\boldsymbol{x}}$, since $\Phi(\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{y}^*(\boldsymbol{x}))$. Also, another non-negative term $\Phi(\boldsymbol{x}) - \Phi^*$ gets smaller as $\boldsymbol{x}$ makes $\Phi(\boldsymbol{x})$ smaller. Thus, as $(\boldsymbol{x};\boldsymbol{y})$ approaches to a minimax optimal point, $V_\lambda(\boldsymbol{x};\boldsymbol{y})$ decreases to near zero. In general, V_λ is not guaranteed to attain exact zero, especially when the objective function $f(\boldsymbol{x};\boldsymbol{y})$ is nonconvex in $\boldsymbol{x}$ (*e.g.*, f is nonconvex-PŁ). Nevertheless, the potential function is still useful for deriving our convergence results.

3.2 Main theorems: upper bounds of convergence rates

Now, we present our main results. We provide a detailed comparison of our theorems against existing results in Section 4. We present the full proof in Appendices C and D. We remark that both Theorems 1 and 2 are special cases (for mini-batch size $b = 1$) of their *mini-batch* extensions: Theorems 4 and 5 in Appendix A.

[1] We say a function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for some $\mu > 0$ if it holds $g(\boldsymbol{x}') \geq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x}\rangle + (\mu/2)\|\boldsymbol{x}' - \boldsymbol{x}\|^2$ $(\forall \boldsymbol{x}, \boldsymbol{x}')$; we say g is μ-strongly concave if $-g$ is μ-strongly convex.

[2] The PŁ(Φ)-PŁ condition is much weaker than **two-sided PŁ** condition assuming "$\boldsymbol{x}$-side" PŁ condition: see Proposition 10. As pointed out by Guo et al. (2020), there exist a PŁ(Φ)-PŁ function $g(\boldsymbol{x};\boldsymbol{y})$ that is not $\boldsymbol{x}$-side μ-PŁ for any $\mu > 0$ but even strongly *concave* in $\boldsymbol{x}$.

Theorem 1 (Nonconvex-PŁ). *Suppose that f satisfies Assumptions 1, 2, 3, and 4. Let $\kappa_2 = L/\mu_2$, where μ_2 is PŁ constant of $-f(\boldsymbol{x};\cdot)$ at all $\boldsymbol{x}$. Let $\lambda = 4$. Choose the step sizes α and β such that*

$$\beta = \min\left\{\frac{1}{6L\sqrt{n(n+A)}},\ \mathcal{O}\left(\left(\frac{V_\lambda(\boldsymbol{z}_0^1)}{Bn^2K}\right)^{\frac{1}{3}}\right)\right\} \quad \textit{and} \quad \alpha = \frac{\beta}{r},$$

for some $r \geq 14\kappa_2^2$. Then, both simSGDA-RR and altSGDA-RR (Algorithm 1) satisfy

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 \leq \mathcal{O}\left(\frac{rLV_\lambda(\boldsymbol{z}_0^1)}{K}\sqrt{1+\frac{A}{n}} + r\left(\frac{L^2BV_\lambda(\boldsymbol{z}_0^1)^2}{nK^2}\right)^{1/3}\right).$$

Upper bound on gradient complexity. To achieve ε-stationarity of the primal function, *i.e.*, $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 \leq \varepsilon^2$, a sufficient number of gradient evaluations (denoted by $T_\varepsilon = nK$) is

$$T_\varepsilon = \mathcal{O}\left(\frac{rLV_\lambda(\boldsymbol{z}_0^1)}{\varepsilon^2}\max\left\{\sqrt{n^2+nA},\ \frac{\sqrt{rnB}}{\varepsilon}\right\}\right).$$

Theorem 2 (PŁ(Φ)-PŁ). *Suppose that f satisfies Assumptions 1, 2, 3, 4, and 5. Let $\kappa_1 = L/\mu_1$ and $\kappa_2 = L/\mu_2$, where μ_1 and μ_2 are PŁ constants of $\Phi(\cdot)$ and $-f(\boldsymbol{x};\cdot)$ (at all $\boldsymbol{x}$), respectively. Let $\lambda = 4$. Choose appropriate step sizes α and β such that*

$$\beta = \min\left\{\frac{1}{6L\sqrt{n(n+A)}},\ \tilde{\mathcal{O}}\left(\frac{\kappa_2^2}{\mu_1 nK}\right)\right\} \quad \textit{and} \quad \alpha = \frac{\beta}{r},$$

for some $r \geq 14\kappa_2^2$. Then, both simSGDA-RR and altSGDA-RR (Algorithm 1) satisfy

$$\mathbb{E}[V_\lambda(\boldsymbol{z}_0^{K+1})] \leq \mathcal{O}\left(V_\lambda(\boldsymbol{z}_0^1)\cdot\exp\left(-\frac{K}{12\kappa_1 r\sqrt{1+\frac{A}{n}}}\right)\right) + \tilde{\mathcal{O}}\left(\frac{\kappa_1^2 r^3 B}{\mu_1 nK^2}\right).$$

Upper bound on gradient complexity. To achieve ε^2-accuracy on expectation of $V_\lambda(\boldsymbol{z}_n^K)$, *i.e.*, $\mathbb{E}[V_\lambda(\boldsymbol{z}_n^K)] \leq \varepsilon^2$, a sufficient number of gradient evaluations (denoted by $T'_\varepsilon = nK$) is

$$T'_\varepsilon = \max\left\{\mathcal{O}\left(\kappa_1 r\sqrt{n^2+nA}\cdot\log\left(\frac{V_\lambda(\boldsymbol{z}_0^1)}{\varepsilon}\right)\right), \tilde{\mathcal{O}}\left(\frac{\kappa_1 r^{3/2}}{\varepsilon}\sqrt{\frac{nB}{\mu_1}}\right)\right\}.$$

Remark on step size ratio. In both theorems, we use the step sizes of ratio $r = \beta/\alpha \gtrsim \kappa_2^2$. It is common to use such a step size scheme $r = \Theta(\kappa_2^2)$ to analyze two-time-scale (S)GDA for nonconvex minimax problems (Jin et al., 2020; Lin et al., 2020; Yang et al., 2020).

Remark on the parameter λ. In our convergence analyses, we arbitrarily choose $\lambda = 4$ which makes the numerical calculations easier. The value of $\lambda > 0$ does not matter for the equivalence between the equation $V_\lambda(\boldsymbol{x}^*;\boldsymbol{y}^*) = 0$ and global minimax condition (Proposition 11). Also, the choice of λ in both theorems can be arbitrary as long as $\lambda > 1$; our logic does not fall apart if other appropriate step sizes for that λ are chosen. That is to say, we can show that the sequence $V_\lambda(\boldsymbol{z}_0^k)$ almost monotonically decreases, ignoring some small variance terms.

4 Comparison with Related Works

4.1 Comparison with stochastic with-replacement setting

First of all, we confirm that SGDA with random reshuffling (RR) has *faster* convergence rates (*i.e.*, fewer gradient computations) than SGDA based on with-replacement sampling. In particular, we compare our results with the analyses on the purely stochastic minimax settings which assume that every stochastic gradient oracle is *independently sampled and unbiased*: this assumption is naturally satisfied by with-replacement sampling for the finite-sum settings we consider. To make the comparisons fair and easy, we simply let $r = \beta/\alpha = \Theta(\kappa_2^2)$, $A = 0$, and $B = \tau^2$.

Lin et al. (2020, Theorem 4.5) present a convergence rate for with-replacement simSGDA with $r = \Theta(\kappa_2^2)$ run on nonconvex μ_2-strongly-concave problems with a convex *bounded* constraint set $\mathcal{Y}$ for dual variable $\boldsymbol{y}$. Their gradient complexity to achieve $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla\Phi(\boldsymbol{x}_t)\right\|^2 \leq \varepsilon^2$ (where T is the

number of iterations) is written as $T_\varepsilon = \mathcal{O}\left(\frac{\kappa_2^2 L\Delta_\Phi + \kappa_2 L^2 D^2}{\varepsilon^2} \max\left\{1, \frac{\kappa_2 \tau^2}{\varepsilon^2}\right\}\right)$, where $\kappa_2 = L/\mu_2$, $\Delta_\Phi = \Phi(\boldsymbol{x}_0) - \Phi^*$, $D = \operatorname{diam} \mathcal{Y}$, and τ^2 is the variance of the (unbiased) stochastic gradient oracles. Their complexity can be simplified as $\mathcal{O}(\kappa_2^3 \tau^2 \varepsilon^{-4})$, treating other factors as constants. In contrast, our Theorem 1 has a better gradient complexity in terms of ε and τ, thanks to shuffling:

$$\mathcal{O}\left(\frac{\kappa_2^2 L V_\lambda(\boldsymbol{z}_0^1)}{\varepsilon^2} \max\left\{n, \frac{\kappa_2 \tau \sqrt{n}}{\varepsilon}\right\}\right), \qquad \text{(Ours, from Theorem 1)}$$

or simply $\mathcal{O}(\kappa_2^3 \tau \sqrt{n} \varepsilon^{-3})$. Thus, our gradient complexity for both simSGDA-RR *and* altSGDA-RR is better than that of with-replacement simSGDA when ε is small as $\varepsilon \le \mathcal{O}(\tau/\sqrt{n})$. Our rate has three more strengths: (i) we do not require strong concavity in $\boldsymbol{y}$, which is a strictly stronger assumption than requiring $\boldsymbol{y}$-side PL condition; (ii) we do not require the constraint set $\mathcal{Y}$ to be bounded; (iii) our result can easily extend to the case of *any* mini-batch sizes, whereas Lin et al. (2020) need a particular choice of mini-batch size $M = \mathcal{O}(\kappa_2 \tau^2/\varepsilon)$ to ensure convergence.

For nonconvex-PŁ objectives, Yang et al. (2022, Theorem 3.1) provide a convergence rate for with-replacement altSGDA with $r = \Theta(\kappa_2^2)$. Their rate can be translated to a gradient complexity for achieving $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left\|\nabla \Phi(\boldsymbol{x}_t)\right\|^2 \le \varepsilon^2$, written as $\mathcal{O}\left(\frac{\kappa_2^2 L V_\lambda(\boldsymbol{z}_0)}{\varepsilon^2}\left(1 + \frac{\kappa_2^2 V_\lambda(\boldsymbol{z}_0)^2 \tau^2}{\Delta_\Phi \varepsilon^2}\right)\right)$ or simply $\mathcal{O}(\kappa_2^4 \tau^2 \varepsilon^{-4})$. Therefore, our gradient complexity for both altSGDA-RR *and* simSGDA-RR is better when ε is small as $\varepsilon \le \mathcal{O}(\kappa_2 \tau/\sqrt{n})$.

For PŁ(Φ)-PŁ objectives, Yang et al. (2020, Theorem 3.3) obtain a convergence rate for with-replacement altSGDA with $r = \Theta(\kappa_2^2)$.[3] They apply diminishing step sizes ($\mathcal{O}(1/t)$, $t \in \mathbb{N}$) to derive a gradient complexity bound $\mathcal{O}\left(\frac{\kappa_1 \kappa_2^4 \tau^2}{\mu_1 \varepsilon^2}\right)$ to achieve $\mathbb{E}[V_\lambda(\boldsymbol{z}_T)] \le \varepsilon^2$. One can apply the constant step sizes depending on the total number T of iterations to their analysis and derive a similar complexity with only deterioration in a logarithmic factor. In contrast, our gradient complexity for both sim/altSGDA-RR using constant step sizes can be written as, for small enough ε,

$$\tilde{\mathcal{O}}\left(\frac{\kappa_1 \kappa_2^3 \tau \sqrt{n}}{\varepsilon \sqrt{\mu_1}}\right). \qquad \text{(Ours, from Theorem 2)}$$

This is a better complexity in ε and κ_2, especially when $\varepsilon \le \tilde{\mathcal{O}}\left(\kappa_2 \tau/\sqrt{n\mu_1}\right)$, even without the requirement of diminishing step size.

4.2 Comparison with other works on stochastic without-replacement setting

One of the most relevant works to this paper is Das et al. (2022, Theorem 3). The authors obtain a similar convergence rate to us for two-sided PŁ objective, based on linearization of gradients, but for a dissimilar algorithm which they refer to as *AGDA-RR*. The algorithm can be also thought of as *epoch-wise*-alternating SGDA-RR, whereas our algorithm (altSGDA-RR) can be called as *step-wise*-alternating SGDA-RR. In epoch k, their algorithm (*i*) performs updates only on $\boldsymbol{x}$ ($\boldsymbol{x}_0^k, \dots, \boldsymbol{x}_n^k$) while fixing $\boldsymbol{y}$ to $\boldsymbol{y}_0^k$, and then (*ii*) performs updates only on $\boldsymbol{y}$ ($\boldsymbol{y}_0^k, \dots, \boldsymbol{y}_n^k$) while fixing $\boldsymbol{x}$ to $\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_n^k$. We believe that our step-wise algorithm is closer to practice, especially when n is large. Because of the distinction between algorithms, the proof techniques are also different.

Xie et al. (2021, Theorem 3) present a convergence rate of *CD-MA*, an extension of simSGDA to the cross-device federated learning setup, on nonconvex-PŁ setting. Their convergence result for *CD-MA* also assumes mini-batch sampling by random reshuffling. As a consequence, they yield a rate analogous to our Theorem 1 if we reduce their result to the single-machine setup. Nevertheless, our convergence bound contains a term that shrinks with the number of components or mini-batches, whereas theirs does not. For a more detailed comparison, please refer to Appendix F.

There are also some works on RR-based (constrained) minimax optimization algorithms other than SGDA, but for convex-concave problems. Maheshwari et al. (2022) present *OGDA-RR*, a gradient-free RR-based optimistic GDA algorithm. Yu et al. (2022) study *stochastic proximal point with RR*, consisting of double-loop epochs. Their analyses exploit convex-concavity and Lipschitz continuity of their objective, based on the arguments by Nagaraj et al. (2019). This enables a direct usage of duality gap, the difference between primal function $\Phi(\cdot)$ and dual function $\Psi(\cdot) = \min_{\boldsymbol{x}} f(\boldsymbol{x}; \cdot)$, as a criterion for optimality. On the contrary, our work relies on a different structure of the functions, which in turn differentiates the constructions of convergence rates.

[3] Although they consider two-sided PŁ problems, their analysis applies to PŁ(Φ)-PŁ problems as well.

4.3 COMPARISON WITH DETERMINISTIC SETTING

Here, we compare our rates with (full-batch two-time-scale) *gradient descent ascent* (GDA):

$$\left[\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{x}_{k-1} - \alpha \nabla_1 f(\boldsymbol{x}_{k-1}; \boldsymbol{y}_{k-1}), \\ \boldsymbol{y}_k &= \boldsymbol{y}_{k-1} + \beta \nabla_2 f(\boldsymbol{x}'; \boldsymbol{y}_{k-1}), \end{aligned}\right. \qquad \text{where } \boldsymbol{x}' = \begin{cases} \boldsymbol{x}_{k-1}, & (\textit{simGDA}), \text{ or} \\ \boldsymbol{x}_k, & (\textit{altGDA}). \end{cases}$$

It uses the whole information of the objective f at every iteration without any noise. For comparison with GDA, we utilize our extended theorems for arbitrary mini-batch size b dividing n (Theorems 4 and 5, Appendix A). By letting $b = n$ and matching our iterate $\boldsymbol{z}_0^k = (\boldsymbol{x}_0^k; \boldsymbol{y}_0^k)$ to a GDA iterate $\boldsymbol{z}_k = (\boldsymbol{x}_k; \boldsymbol{y}_k)$, our results reduce to upper convergence bounds for simGDA and altGDA.

For nonconvex-PŁ problems (Theorems 1 & 4), the convergence rate and *iteration* complexity (*i.e.*, sufficient number of iterations K_ε) become

$$\min_{k \in [K]} \|\nabla \Phi(\boldsymbol{x}_k)\|^2 \le \mathcal{O}\left(\frac{\kappa_2^2 L V_\lambda(\boldsymbol{z}_1)}{K}\right); \quad \textit{i.e.},\ K_\varepsilon = \mathcal{O}\left(\frac{\kappa_2^2 L V_\lambda(\boldsymbol{z}_1)}{\varepsilon^2}\right), \tag{3}$$

when $r = \Theta(\kappa_2^2)$. This is similar to a known rate of simGDA with $r = \Theta(\kappa_2^2)$ for nonconvex-strongly-concave problems by Lin et al. (2020, Theorem 4.4) as a special case. Their iteration complexity is written as $\mathcal{O}((\kappa_2^2 L \Delta_\Phi + \kappa_2 L^2 D^2)/\varepsilon^2)$, where the symbols are already defined in Section 4.1. To see how the two bounds compare in terms of the factors other than ε, notice that we have $\Phi(\boldsymbol{x}) - f(\boldsymbol{x}; \boldsymbol{y}) \le \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{y}^*(\boldsymbol{x})\|^2$ for any $(\boldsymbol{x}; \boldsymbol{y})$, due to L-smoothness of $-f$. Here, $\boldsymbol{y}^*(\boldsymbol{x})$ is an element of $\mathcal{Y}_{\boldsymbol{x}}^* = \arg\max_y f(\boldsymbol{x}; \boldsymbol{y})$. Thus, we have $V_\lambda(\boldsymbol{z}_1) = \lambda[\Phi(\boldsymbol{x}_1) - \Phi^*] + [\Phi(\boldsymbol{x}_1) - f(\boldsymbol{z}_1)] \le \lambda \Delta_\Phi + LD^2/2$. As a result, we could *loosely* translate our iteration complexity (3) to $\mathcal{O}((\kappa_2^2 L \Delta_\Phi + \kappa_2^2 L^2 D^2)/\varepsilon^2)$. We suspect that the discrepancy in terms of κ_2 comes from the fact that our analysis does not require the (strong) concavity in terms of $\boldsymbol{y}$ or a bounded constraint $\mathcal{Y}$: these requirements made a considerable difference in proofs.

For PŁ(Φ)-PŁ problems (Theorems 2 & 5), the rate and iteration complexity (K'_ε) become

$$V_\lambda(\boldsymbol{z}_{K+1}) \le V_\lambda(\boldsymbol{z}_1) \cdot \exp\left(-\frac{K}{C \kappa_1 \kappa_2^2}\right); \quad \textit{i.e.},\ K'_\varepsilon = \mathcal{O}\left(\kappa_1 \kappa_2^2 \log(1/\varepsilon)\right) \tag{4}$$

where $r = \Theta(\kappa_2^2)$ and C is a numerical constant. This recovers the linear convergence by Yang et al. (2020, Theorem 3.2) as a special case, where they prove convergence of altGDA with step size ratio $r = \Theta(\kappa_2^2)$ for two-sided PŁ problem. Following the proof of (Yang et al., 2020, Theorem 3.2), one can show that the bound (4) indeed implies the actual convergence to a global minimax point $\boldsymbol{z}^*$, in the sense that we can achieve $\|\boldsymbol{z}_k - \boldsymbol{z}^*\| \le \varepsilon$ in $\mathcal{O}\left(\kappa_1 \kappa_2^2 \log(1/\varepsilon)\right)$ iterations.

5 LOWER BOUND FOR TWO-TIME-SCALE SIMGDA

As an extension of the discussion from Section 4.3, we characterize a lower complexity bound of two-time-scale simGDA with arbitrary step size ratio $r = \beta/\alpha$ for smooth strongly-convex-strongly-concave (SC-SC) cases. At least for $r \gtrsim \kappa_2^2$, our lower bound matches the upper complexity bound of GDA for a much wider class of problems: smooth PŁ(Φ)-PŁ problems,[4] which is quite surprising.

For a smooth PŁ(Φ)-PŁ problem, simGDA with step size ratio at least $r = \Omega(\kappa_2^2)$ has an upper complexity bound $K = \mathcal{O}(\kappa_1 r \log(1/\varepsilon))$ for a *global* ε-convergence $V_\lambda(\boldsymbol{z}_K) \le \varepsilon^2$ in terms of potential function. This means that the lowest complexity is $\mathcal{O}(\kappa_1 \kappa_2^2 \log(1/\epsilon))$, achieved when $r = \Theta(\kappa_2^2)$. On the other hand, for a L-smooth μ-SC-SC problem with saddle point $\boldsymbol{z}^*$, it is well-known that the one-time-scale ($\alpha = \beta$) simGDA has a tight upper/lower complexity $K = \Theta(\kappa^2 \log(1/\varepsilon))$ to achieve $\|\boldsymbol{z}_K - \boldsymbol{z}^*\|^2 \le \varepsilon^2$, where $\kappa = L/\mu$ (*e.g.*, Das et al. (2022, Theorem C.1)). The difference of complexity bounds in condition number ($\kappa_1 \kappa_2^2$ v.s. κ^2) is somewhat questionable because, at least in smooth minimization problems, strongly convex problems and PŁ problems have identical gradient descent (GD) iteration complexity $\mathcal{O}(\kappa \log(1/\varepsilon))$ (Karimi et al., 2016, Theorem 1).

One could ask where the discrepancy in terms of κ comes from: is it due to (*i*) the criteria ($V_\lambda(\boldsymbol{z}_K)$ v.s. $\|\boldsymbol{z}_K - \boldsymbol{z}^*\|^2$) for ε-accuracy, (*ii*) the function classes (PŁ(Φ)-PŁ v.s. SC-SC), or (*iii*) the step size

[4] strongly-convex-strongly-concave $\subset$ two-sided PŁ $\subset$ PŁ(Φ)-PŁ $\subset$ nonconvex-PŁ.

ratios ($\Omega(\kappa_2^2)$ v.s. 1)? We answer the question by showing the following theorem: the discrepancy in κ comes from the step size ratio difference. We defer the proof to Appendix E.

Theorem 3 (Lower bound, ratio-specific)**.** *Consider a class $\mathcal{F}(L, \mu_1, \mu_2)$ of functions $f(\boldsymbol{x}; \boldsymbol{y})$ with two arguments $\boldsymbol{x}$ and $\boldsymbol{y}$, which is L-smooth, μ_1-strongly-convex in $\boldsymbol{x}$, and μ_2-strongly-concave in $\boldsymbol{y}$. Suppose $\kappa_1 = L/\mu_1 \geq c$ and $\kappa_2 = L/\mu_2 \geq c$ for a constant $c > 1$. Then, for any step size ratio $r = \beta/\alpha > 0$, there exists a function $f \in \mathcal{F}(L, \mu_1, \mu_2)$ with a unique saddle point $\boldsymbol{z}^*$, for which two-time-scale simGDA with any step sizes $(\alpha, \beta) = (\beta/r, \beta)$ requires at least*

$$K = \begin{cases} \Omega\left(\kappa_1 r \log(1/\varepsilon)\right), & \text{if } r \geq \kappa_2/c, \\ \Omega\left(\kappa_1 \kappa_2 \log(1/\varepsilon)\right), & \text{if } c/\kappa_1 \leq r \leq \kappa_2/c, \\ \Omega((\kappa_2/r)\log(1/\varepsilon)), & \text{if } 0 < r \leq c/\kappa_1 \end{cases}$$

iterations to achieve either $\|\boldsymbol{z}_K - \boldsymbol{z}^*\|^2 \leq \varepsilon^2$ or $V_\lambda(\boldsymbol{z}_K) \leq \varepsilon^2$.

Thanks to Theorem 3, we can say from Theorem 5 that for any step size ratio $r \gtrsim \kappa_2^2$, we have a *tight* upper bound on the iteration complexity $K = \mathcal{O}(\kappa_1 r \log(1/\varepsilon))$ of two-time-scale simGDA for general PŁ(Φ)-PŁ problems. Note that Theorem 3 also subsumes the existing lower bound of the one-time-scale ($r = 1$) GDA for SC-SC problems.

Given the tightness of bounds for $r \gtrsim \kappa_2^2$, a natural next step is to discuss $1 \lesssim r \lesssim \kappa_2^2$. A recent work by Li et al. (2022) also discusses the step size ratio of simGDA. In Li et al. (2022, Theorem 4.1), the authors construct a y-side strongly-concave function[5] and show that simGDA with a step size ratio $r \leq \kappa_2$ is *impossible* to converge. The *necessity* of $r \gtrsim \kappa_2$ implied by this theorem also applies to the PŁ(Φ)-PŁ case. Thus, there is no hope for showing an upper convergence bound of simGDA with $1 \lesssim r \lesssim \kappa_2$ for general nonconvex-PŁ problems. We remark that their theorem does not contradict nor subsume Theorem 3 because we consider a much smaller function class (SC-SC) to construct the lower bounds.

On the *sufficiency* of $r \gtrsim \kappa_2$ for convergence, Li et al. (2022, Theorem 4.2) show that simGDA with $r \geq c\kappa$ (for some $c > 1$) can *locally* converge at the iteration complexity $\mathcal{O}(\kappa_1 r \log(1/\varepsilon))$ for some nonconvex-strongly-concave problems, which matches the bound in Theorem 3. Our upper bounds (Theorems 4 and 5) do require $r \gtrsim \kappa_2^2$, which may look suboptimal, but we claim that our results are not necessarily weaker. One reason is that our convergence guarantee is *global*, *i.e.*, independent of the initialization. Another reason is that their analysis is only valid when a *differential Stackelberg equilibrium*[6] exists, whereas a general PŁ(Φ)-PŁ function may not have such an equilibrium (for an example, see Proposition 13 in Appendix B).

As far as we know, it is still an open problem whether a *global* convergence bound for simGDA on nonconvex-PŁ problems can be shown when the step size ratio r is between $\Omega(\kappa_2)$ and $\mathcal{O}(\kappa_2^2)$.

6 Experiments

To validate our theoretical findings, here we present some numerical results. We focus on the primal-PŁ-strongly-concave (or PŁ(Φ)-SC, which is PŁ(Φ)-PŁ as well) quadratic games of the form

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^d} \max_{\boldsymbol{y}\in\mathbb{R}^d} f(\boldsymbol{x};\boldsymbol{y}) = \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y} = \tfrac{1}{n}\textstyle\sum_{i=1}^n f_i(\boldsymbol{x};\boldsymbol{y}), \\ &\text{where} \quad f_i(\boldsymbol{x};\boldsymbol{y}) = \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{B}_i\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{y}^\top \boldsymbol{C}_i\boldsymbol{y} + \boldsymbol{u}_i^\top \boldsymbol{x} - \boldsymbol{v}_i^\top \boldsymbol{y}. \end{aligned} \tag{5}$$

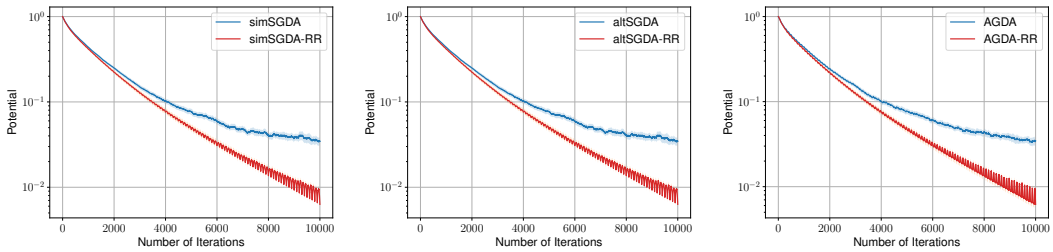
This toy example is often used to numerically evaluate the minimax algorithms (Yang et al., 2020; Loizou et al., 2021; Das et al., 2022) and appears in various domains such as AUC maximization (Ying et al., 2016), policy evaluation (Du et al., 2017), and imitation learning (Cai et al., 2019)

To make the game in Equation (5) satisfy PŁ(Φ)-SC and component L-smoothness, we need to sample the matrices and vectors carefully. First, they have to satisfy $\|\boldsymbol{A}_i\|_2, \|\boldsymbol{B}_i\|_2, \|\boldsymbol{C}_i\|_2 \leq L$, $\sum_{i=1}^n \boldsymbol{u}_i = \sum_{i=1}^n \boldsymbol{v}_i = \boldsymbol{0}$. To make the primal function Φ a well-defined real-valued function for any $\boldsymbol{x} \in \mathbb{R}^d$, we choose $\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{C}_i$ to be positive definite, in particular $\mu\boldsymbol{I} \preceq \boldsymbol{C}$ where $\boldsymbol{I}$ is an identity matrix. Then, the primal function can be explicitly written as

$$\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathbb{R}^d} f(\boldsymbol{x};\boldsymbol{y}) = \tfrac{1}{2}\boldsymbol{x}^\top \left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top\right)\boldsymbol{x} := \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}.$$

[5] $g(x;y) = -\frac{L}{2}x^2 + Lxy - \frac{\mu}{2}y^2$, where $L/\mu > 1$: its primal function is strongly convex.

[6] Loosely speaking, a differential Stackelberg equilibrium is a stationary point $(\boldsymbol{x}^*; \boldsymbol{y}^*)$ where $f(\boldsymbol{x}^*; \cdot)$ is locally strongly concave near $\boldsymbol{y}^*$ and $\Phi(\cdot)$ is locally strongly convex near $\boldsymbol{x}^*$.



(a) simSGDA v.s. simSGDA-RR. (b) altSGDA v.s. altSGDA-RR. (c) AGDA v.s. AGDA-RR.

Figure 1: Experimental results on quadratic games (5). The solid lines are the average across 50 different runs and 95% confidence intervals (± 1.96 std) are shown as the shaded region. The values are normalized by dividing them by the initial value. The vertical axes are on a logarithmic scale. The figure clearly shows better performance of algorithms using RR.

We choose the matrix $\boldsymbol{M} := \boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top$ to be rank-deficient positive semi-definite. By letting the smallest nonzero eigenvalue of $\boldsymbol{M}$ be μ, we ensure that Φ is μ-PŁ, but not strongly convex since $\boldsymbol{M}$ is not of full rank. We emphasize that the objective function f is not even (strongly-)convex in $\boldsymbol{x}$ in general. We compare six algorithms in total: simSGDA-RR, altSGDA-RR, AGDA-RR (as defined in Das et al. (2022)), and the with-replacement counterparts of these three algorithms. To this end, on 50 different randomly-generated quadratic games, we run each algorithm for the same number of epochs using constant step sizes. Lastly, we use the constant step sizes of ratio $\beta/\alpha = c\kappa^2$ for some constant c and $\kappa = L/\mu$. Refer to Appendix G for further details on the parameter choices.

We report the potential function values (V_λ, defined in Equation (2)) at every iteration.[7] Results are presented in Figure 1. As we discussed in Section 4.1, we observe that the random reshuffling considerably accelerates the convergence of the algorithms. Furthermore, all three algorithms with random reshuffling show more or less the same performance. Specifically, the plots for simSGDA(resp. simSGDA-RR) and altSGDA(resp. altSGDA-RR) are almost identical (the numerical differences per iteration is roughly $\approx 10^{-2}$). We believe this is because we choose a random seed for each of the 50 different runs and share it across different algorithms.

Here are a couple of additional observations. First, random reshuffling is more advantageous when the discrepancy between component functions is larger (*i.e.*, larger A and B in Assumption 2). We control the discrepancy by manipulating the maximum absolute entry of $\boldsymbol{u}_i$'s and $\boldsymbol{v}_i$'s. Second, the larger $\kappa = L/\mu$ is, the more iterations are necessary for the convergence of every algorithm we observe. For more discussions on experimental results, please refer to Appendix G.

7 Conclusion

We investigated two-time-scale stochastic algorithms based on without-replacement component sampling, called simSGDA-RR and altSGDA-RR, for solving smooth nonconvex finite-sum minimax optimization problems. We established convergence rates under the $\boldsymbol{y}$-side PŁ condition (nonconvex-PŁ) and, additionally, the primal PŁ condition (PŁ(Φ)-PŁ). We ascertain that the SGDA-RR can achieve a faster rate than its with-replacement counterpart, which agrees with the existing theory on without-replacement SGD for minimization. Lastly, we provided lower bounds for two-time-scale simGDA with any fixed step size ratio r, demonstrating that the full-batch upper bound with $r \gtrsim \kappa_2^2$ for PŁ(Φ)-PŁ functions is tight. Possible future directions include widening our results beyond sim/altSGDA (*e.g.*, extra-gradient or optimistic GDA) and beyond RR (*e.g.*, single/adversarial shuffling). As also discussed in Section 5, an interesting open question remains open: can we identify tight convergence rates for stochastic (with-/without-replacement) and/or deterministic GDA with step size ratio r satisfying $\kappa_2 \lesssim r \lesssim \kappa_2^2$, for general nonconvex-PŁ problems?

[7] As described in Section 4.2, AGDA-RR uses only one-side gradient (∇_1 or ∇_2) at each iteration; given a fixed budget of gradient computations, it should access components twice as many times as SGDA-RR. Hence, we report the values at every other iteration of AGDA & AGDA-RR, for a fair comparison.

REFERENCES

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33: 17526–17535, 2020.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. *arXiv preprint arXiv:2202.07262*, 2022.

Qi Cai, Mingyi Hong, Yongxin Chen, and Zhaoran Wang. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*, 2019.

Aniket Das, Bernhard Schölkopf, and Michael Muehlebach. Sampling without replacement leads to faster rates in finite-sum minimax optimization. *arXiv preprint arXiv:2206.02953*, 2022.

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058. PMLR, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Zhishuai Guo, Zhuoning Yuan, Yan Yan, and Tianbao Yang. Fast objective & duality gap convergence for nonconvex-strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, England, 2 edition, October 2012.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Haochuan Li, Farzan Farnia, Subhro Das, and Ali Jadbabaie. On convergence of gradient descent ascent: A tight local analysis. In *International Conference on Machine Learning*, pp. 12717–12740. PMLR, 2022.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4213–4220, 2019.

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

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic AUC maximization with deep neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJepXaVYDr.

Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34: 19095–19108, 2021.

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

Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, Shankar Sastry, and Lillian Ratliff. Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 6702–6734. PMLR, 2022.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning (ICML)*, pp. 4703–4711. PMLR, 2019.

Lam M Nguyen, Quoc Tran-Dinh, Dzung T Phan, Phuong Ha Nguyen, and Marten Van Dijk. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

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

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, pp. 7964–7973. PMLR, 2020.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

Jiahao Xie, Chao Zhang, Zebang Shen, Weijie Liu, and Hui Qian. Efficient cross-device federated learning algorithms for minimax problems. *arXiv e-prints*, pp. arXiv–2105, 2021.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800, 2020.

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

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

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online AUC maximization. *Advances in neural information processing systems*, 29, 2016.

Yaodong Yu, Tianyi Lin, Eric V Mazumdar, and Michael Jordan. Fast distributionally robust learning with variance-reduced min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1219–1250. PMLR, 2022.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3040–3049, 2021.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Open problem: Can single-shuffle SGD be better than reshuffling SGD and GD? In *Conference on Learning Theory*, pp. 4653–4658. PMLR, 2021.

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

CONTENTS

1 Introduction **1**

1.1 Summary of our contributions 2

2 Problem setup **2**

2.1 Notation 2

2.2 Algorithms: (two-time-scale) simSGDA-RR & altSGDA-RR 3

2.3 Assumptions and definitions 3

3 Main results **4**

3.1 Potential function V_λ 4

3.2 Main theorems: upper bounds of convergence rates 4

4 Comparison with related works **5**

4.1 Comparison with stochastic with-replacement setting 5

4.2 Comparison with other works on stochastic without-replacement setting 6

4.3 Comparison with deterministic setting 7

5 Lower bound for two-time-scale simGDA **7**

6 Experiments **8**

7 Conclusion **9**

A Mini-batch SGDA-RR and convergence rates **14**

B Technical propositions **16**

B.1 Function classes: PŁ condition, smoothness, and more 16

B.2 Without-replacement sampling 20

B.3 Basic recurrence inequality 22

C Proofs for (mini-batch) simultaneous SGDA with random reshuffling **23**

C.1 Warm-up: proof sketch for $b = 1$ 23

C.2 Epoch-wise representations and bounding noise terms 24

C.3 Recurrence inequalities for general smooth nonconvex-PŁ objective 28

C.4 Convergence rates for smooth nonconvex-PŁ problem 30

C.5 Convergence rates for smooth PŁ(Φ)-PŁ problem 31

D Proofs for (mini-batch) alternating SGDA-RR: focusing on changes in the proof **34**

D.1 Epoch-wise representations and bounding noise terms 34

D.2 Bounding noise terms: a bit different proof of Lemma 18 34

D.3 Recurrence inequalities for general smooth nonconvex-PŁ objective 35

D.4 Small step size assumptions 36

E Proofs for lower bound of two-time-scale GDA **37**

F Omitted comparison with related works **42**

F.1 Comparison with Xie et al. (2021) 42

G Description on Experiments **43**

G.1 Parameter choices 43

G.2 Comparison: the effect of component discrepancy 44

G.3 Comparison: the effect of condition number 44

A MINI-BATCH SGDA-RR AND CONVERGENCE RATES

Here, we present an algorithm that extends simSGDA-RR and altSGDA-RR by using mini-batches of size $b \geq 1$. For simplicity, we assume that the number of components n is an integer multiple of the mini-batch size b for simplicity; *i.e.*, for some integer $q \geq 1$, $n = bq$. One can easily extend this to the case when n does not need to be a multiple of b (namely, $n = b(q-1) + s$, where $q \geq 1$, $s \in [b]$) so that there are $q-1$ mini-batches of size b and one more mini-batch of size $s \leq b$.

Algorithm 2 Mini-batch simSGDA/altSGDA-RR

Given: The number of components n; the size of a mini-batch b (a divisor of n); the number of epochs K; step sizes $\alpha, \beta > 0$
Initialize: $(\boldsymbol{x}_0^1; \boldsymbol{y}_0^1) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$
for $k \in [K]$ **do**
  Sample $\sigma_k \sim \mathrm{Unif}(\mathbb{S}_n)$ ▷ RR: uniformly randomly shuffle the indices every epoch
  for $t \in [n/b]$ **do**
    $\mathcal{B}_t^k := \{\sigma_k(b(t-1)+1), \cdots, \sigma_k(bt)\}$ ▷ Mini-batch : a set of component indices
    $\boldsymbol{x}_t^k = \boldsymbol{x}_{t-1}^k - \frac{\alpha}{b} \sum_{i \in \mathcal{B}_t^k} \nabla_1 f_i(\boldsymbol{x}_{t-1}^k; \boldsymbol{y}_{t-1}^k)$
    if simSGDA-RR **then**
      $\boldsymbol{y}_t^k = \boldsymbol{y}_{t-1}^k + \frac{\beta}{b} \sum_{i \in \mathcal{B}_t^k} \nabla_2 f_i(\boldsymbol{x}_{t-1}^k; \boldsymbol{y}_{t-1}^k)$ ▷ simultaneous update: $\boldsymbol{x}$ & $\boldsymbol{y}$
    else if altSGDA-RR **then**
      $\boldsymbol{y}_t^k = \boldsymbol{y}_{t-1}^k + \frac{\beta}{b} \sum_{i \in \mathcal{B}_t^k} \nabla_2 f_i(\boldsymbol{x}_t^k; \boldsymbol{y}_{t-1}^k)$ ▷ alternating update: $\boldsymbol{x} \to \boldsymbol{y}$
  $(\boldsymbol{x}_0^{k+1}; \boldsymbol{y}_0^{k+1}) = (\boldsymbol{x}_{n/b}^k; \boldsymbol{y}_{n/b}^k)$

Next, we illustrate the generalized versions of our main results (Theorems 1 and 2) for Algorithm 2 with mini-batches of size $b \geq 1$. Let us assume $n \geq 2$ because the case $n = 1$ trivially boils down to simGDA or altGDA. We defer the proofs for simultaneous updates to Appendix C. We present the parts that change in the proof for alternating updates in Appendix D.

Theorem 4 (Nonconvex-PŁ, mini-batch SGDA-RR)**.** *Suppose that f satisfies Assumptions 1, 2, 3, and 4. Let $\lambda = 4$. Choose the step sizes α and β by $\alpha = \beta/r$ for some $r \geq 14\kappa_2^2$ and*

$$\beta = b \cdot \min\left\{ \frac{1}{6Ln\sqrt{1 + \frac{n-b}{n-1} \cdot \frac{A}{n}}}, \ \frac{1}{c}\left(\frac{V_\lambda(\boldsymbol{z}_0^1)}{Ln^2(\frac{n-b}{n-1})BK} \right)^{\frac{1}{3}} \right\},$$

for some numerical constant $c > 0$. Then, mini-batch simSGDA-RR and altSGDA-RR with mini-batch size b (a divisor of n) satisfy

$$\begin{aligned} \min_{k \in [K]} \mathbb{E}\left\|\nabla \Phi(\boldsymbol{x}_0^k)\right\|^2 &\leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left\|\nabla \Phi(\boldsymbol{x}_0^k)\right\|^2 \\ &\leq \frac{6rLV_\lambda(\boldsymbol{z}_0^1)}{K} \sqrt{1 + \left(\frac{n-b}{n-1}\right)\frac{A}{n}} + 2cr\left(\frac{L^2 B\, V_\lambda(\boldsymbol{z}_0^1)^2}{nK^2} \cdot \frac{n-b}{n-1} \right)^{1/3}. \end{aligned}$$

Theorem 5 (PŁ(Φ)-PŁ, mini-batch SGDA-RR)**.** *Suppose that f satisfies Assumptions 1, 2, 3, 4, and 5. Let $\lambda = 4$. Choose the step sizes α and β by $\alpha = \beta/r$ for some $r \geq 14\kappa_2^2$ and*

$$\beta = b \cdot \min\left\{ \frac{1}{6Ln\sqrt{1 + \frac{n-b}{n-1} \cdot \frac{A}{n}}}, \ \frac{2r}{\mu_1 nK} \max\left\{ 1, \ \log\left(\frac{V_\lambda(\boldsymbol{z}_0^1)\mu_1 nK^2}{8c^3\kappa_1^2 r^3 \left(\frac{n-b}{n-1}\right) B} \right) \right\} \right\},$$

for some numerical constant $c > 0$. Then, mini-batch simSGDA-RR and altSGDA-RR with mini-batch size b (a divisor of n) satisfy

$$\mathbb{E}[V_\lambda(\boldsymbol{z}_0^{K+1})] \leq \mathcal{O}\left(V_\lambda(\boldsymbol{z}_0^1) \cdot \exp\left(-\frac{K}{12\kappa_1 r\sqrt{1 + \frac{n-b}{n-1}\frac{A}{n}}} \right) \right) + \tilde{\mathcal{O}}\left(\frac{\kappa_1^2 r^3 B}{\mu_1 nK^2} \right) \cdot \frac{n-b}{n-1}.$$

As a side remark, some works consider a sampling method called *b-minibatch sampling* where all the elements in each mini-batch are distinct (*i.e.*, without-replacement component sampling per mini-batch), *e.g.*, Loizou et al. (2021, Definition 2.1). However, there is a significant gap between this method and ours: any two distinct mini-batches sampled by the b-minibatch sampling can intersect to each other (*i.e.*, mini-batches are sampled with replacement), whereas, in each epoch of our Algorithm 2, all the mini-batches are mutually disjoint.

B TECHNICAL PROPOSITIONS

Notation. Throughout this appendix, we use $\mathcal{X} = \mathbb{R}^{d_x}$ and $\mathcal{Y} = \mathbb{R}^{d_y}$. Given a closed set $\mathcal{S} \subset \mathbb{R}^d$, we denote the set of all projection(s) of $\boldsymbol{v} \in \mathbb{R}^d$ onto $\mathcal{S}$, *i.e.*, the nearest point(s) in $\mathcal{S}$ from $\boldsymbol{v}$, by $\Pi_{\mathcal{S}}(\boldsymbol{v}) := \arg\min_{\boldsymbol{w}\in\mathcal{S}} \|\boldsymbol{v} - \boldsymbol{w}\|$.

B.1 FUNCTION CLASSES: PŁ CONDITION, SMOOTHNESS, AND MORE

Proposition 6 ($\kappa \geq 1$)**.** *Let g be an L-smooth function which is bounded below by g^*. Then, for any $\boldsymbol{x}$,*

$$\|\nabla g(\boldsymbol{x})\|^2 \leq 2L\left[g(\boldsymbol{x}) - g^*\right].$$

If g is μ-PŁ as well, then $\mu \leq L$. Consequently, the condition number $\kappa := L/\mu$ of g is ≥ 1.

Proof. Since g is L-smooth, for any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$g^* \leq g(\boldsymbol{y}) \leq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{6}$$

Now define a convex quadratic function $h_x(\boldsymbol{y})$ of $\boldsymbol{y}$ as

$$h_x(\boldsymbol{y}) := g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Since its gradient is

$$\nabla h_x(\boldsymbol{y}) = \nabla g(\boldsymbol{x}) + L(\boldsymbol{y} - \boldsymbol{x}),$$

$\boldsymbol{y}^* := \boldsymbol{x} - \frac{1}{L}\nabla g(\boldsymbol{x})$ is a minimum of h_x. Plugging $\boldsymbol{y} = \boldsymbol{y}^*$ to the equation (6), we get

$$g^* \leq g(\boldsymbol{x}) + \left\langle \nabla g(\boldsymbol{x}), -\frac{1}{L}\nabla g(\boldsymbol{x})\right\rangle + \frac{L}{2}\left\|-\frac{1}{L}\nabla g(\boldsymbol{x})\right\|^2 = g(\boldsymbol{x}) - \frac{1}{2L}\|\nabla g(\boldsymbol{x})\|^2.$$

Rearranging the terms,

$$\|\nabla g(\boldsymbol{x})\|^2 \leq 2L\left[g(\boldsymbol{x}) - g^*\right].$$

If we additionally utilize PŁ inequality with $g^* := \min g(\boldsymbol{x})$,

$$\|\nabla g(\boldsymbol{x})\|^2 \geq 2\mu\left[g(\boldsymbol{x}) - g^*\right],$$

we directly yield $\mu \leq L$ and thus $\kappa = L/\mu \geq 1$. □

Definition 1 (Karimi et al. (2016))**.** *Consider $g : \mathcal{X} \to \mathbb{R}$. Let $\boldsymbol{x}_p \in \Pi_{\mathcal{X}^*}(\boldsymbol{x})$ be a projection of $\boldsymbol{x}$ onto the optimal set $\mathcal{X}^* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x})$.*

(1) We say g satisfies μ-strong convexity (SC) if $g(\boldsymbol{x}') \geq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x}\rangle + \frac{\mu}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|^2$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.

(2) We say g satisfies μ-restricted secant inequality (RSI) if $\langle \nabla g(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_p\rangle \geq \mu\|\boldsymbol{x}_p - \boldsymbol{x}\|^2$ for any $\boldsymbol{x} \in \mathcal{X}$.

(3) We say g satisfies μ-error bound (EB) condition if $\|\nabla g(\boldsymbol{x})\| \geq \mu\|\boldsymbol{x}_p - \boldsymbol{x}\|$ for any $\boldsymbol{x} \in \mathcal{X}$.

(4) We say g satisfies μ-quadratic growth (QG) condition if $g(\boldsymbol{x}) - \min_{\boldsymbol{x}'} g(\boldsymbol{x}') \geq \frac{\mu}{2}\|\boldsymbol{x}_p - \boldsymbol{x}\|^2$ for any $\boldsymbol{x} \in \mathcal{X}$.

Proposition 7. *From Definition 1, The following implications are true.*

- *μ-SC implies μ-PŁ and μ-RSI.*
- *μ-PŁ implies μ-QG and μ-EB.*
- *μ-RSI implies μ-EB.*

- *μ-EB and L-smoothness together imply (μ^2/L)-PŁ.*

Proof. Most of the proofs originated from Karimi et al. (2016, Theorem 2).

(SC $\Rightarrow$ PŁ) Substitute $\boldsymbol{x}$ to $\boldsymbol{x}_p$ and $\boldsymbol{x}'$ to $\boldsymbol{x}$, respectively, from Definition 1.(1).

(PŁ $\Rightarrow$ QG & EB) See the proof in Karimi et al. (2016, Theorem 2)

(SC $\Rightarrow$ RSI) We know μ-SC $\Rightarrow$ μ-PŁ $\Rightarrow$ μ-QG. From Definition 1.(1) & 1.(4),

$$\begin{aligned}\langle \nabla g(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_p \rangle &\overset{\text{SC}}{\geq} g(\boldsymbol{x}) - g(\boldsymbol{x}_p) + \frac{\mu}{2} \|\boldsymbol{x}_p - \boldsymbol{x}\|^2 \\ &\overset{\text{QG}}{\geq} \frac{\mu}{2} \|\boldsymbol{x}_p - \boldsymbol{x}\|^2 + \frac{\mu}{2} \|\boldsymbol{x}_p - \boldsymbol{x}\|^2 = \mu \|\boldsymbol{x}_p - \boldsymbol{x}\|^2 .\end{aligned}$$

This implies μ-RSI.

(RSI $\Rightarrow$ EB) See the proof in Karimi et al. (2016, Theorem 2).

(EB & smooth $\Rightarrow$ PŁ) We use $\nabla g(\boldsymbol{x}_p) = \mathbf{0}$. By L-smoothness and μ-EB condition,

$$\begin{aligned} g(\boldsymbol{x}) - g(\boldsymbol{x}_p) &\overset{\text{smooth}}{\leq} \langle \nabla g(\boldsymbol{x}_p), \boldsymbol{x} - \boldsymbol{x}_p \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_p\|^2 = \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_p\|^2 \\ &\overset{\text{EB}}{\leq} \frac{L}{2\mu^2} \|\nabla g(\boldsymbol{x})\|^2 .\end{aligned}$$

This implies (μ^2/L)-PŁ condition on g. □

Proposition 8 (Lipschitz continuity-like property of $\boldsymbol{y}^*(\boldsymbol{x})$)**.** *For an L-smooth function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, suppose $-g(\boldsymbol{x}; \cdot)$ is μ_2-PŁ. Let $\kappa_2 = L/\mu_2$.*

Consider any $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathcal{X}$. For any $\boldsymbol{y}_0^* \in \mathcal{Y}_{\boldsymbol{x}_0}^* = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} g(\boldsymbol{x}_0; \boldsymbol{y})$, there exists a $\boldsymbol{y}_1^* \in \mathcal{Y}_{\boldsymbol{x}_1}^* = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} g(\boldsymbol{x}_1; \boldsymbol{y})$ such that $\|\boldsymbol{y}_0^* - \boldsymbol{y}_1^*\| \leq \kappa_2 \|\boldsymbol{x}_0 - \boldsymbol{x}_1\|$.

In fact, it is enough to choose $\boldsymbol{y}_1^*$ as a projection of $\boldsymbol{y}_0^*$ onto the set $\mathcal{Y}_{\boldsymbol{x}_1}^*$, namely, $\boldsymbol{y}_1^* \in \Pi_{\mathcal{Y}_{\boldsymbol{x}_1}^*}(\boldsymbol{y}_0^*)$.

Proof. We borrow the proof from Nouiehed et al. (2019, Lemma A.3).

Recall $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}' \in \mathcal{Y}} g(\boldsymbol{x}; \boldsymbol{y}')$. By PŁ inequality and smoothness of g,

$$\begin{aligned} 2\mu_2 \left(\Phi(\boldsymbol{x}_1) - g(\boldsymbol{x}_1; \boldsymbol{y}_0^*)\right) &\leq \|\nabla_2 g(\boldsymbol{x}_1; \boldsymbol{y}_0^*)\|^2 \\ &= \|\nabla_2 g(\boldsymbol{x}_1; \boldsymbol{y}_0^*) - \nabla_2 g(\boldsymbol{x}_0; \boldsymbol{y}_0^*)\|^2 \leq L^2 \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 .\end{aligned}$$

The second equality applies $\nabla_2 g(\boldsymbol{x}_0; \boldsymbol{y}_0^*) = \mathbf{0}$, since $\boldsymbol{y}_0^* \in \arg\max_{\boldsymbol{y}} g(\boldsymbol{x}_0; \boldsymbol{y})$.

Moreover, note that $-g(\boldsymbol{x}_1; \cdot)$ satisfies μ_2-QG condition ($\because$ Proposition 7). To apply this, we utilize our choice of $\boldsymbol{y}_1^*$:

$$\Phi(\boldsymbol{x}_1) - g(\boldsymbol{x}_1; \boldsymbol{y}_0^*) \geq \frac{\mu_2}{2} \|\boldsymbol{y}_1^* - \boldsymbol{y}_0^*\|^2 .$$

As a result, we have $\mu_2^2 \|\boldsymbol{y}_0^* - \boldsymbol{y}_1^*\|^2 \leq L^2 \|\boldsymbol{x}_0 - \boldsymbol{x}_1\|^2$. This completes the proof. □

Proposition 9 (Smoothness of primal function)**.** *Consider the same function g as Proposition 8. Then, the function $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}' \in \mathcal{Y}} g(\boldsymbol{x}; \boldsymbol{y}')$ is differentiable with*

$$\nabla \Phi(\boldsymbol{x}) = \nabla_1 g(\boldsymbol{x}; \boldsymbol{y}^*(\boldsymbol{x})), \quad \textit{regardless of the choice of } \boldsymbol{y}^*(\boldsymbol{x}) \in \arg\max_{\boldsymbol{y}' \in \mathcal{Y}} g(\boldsymbol{x}; \boldsymbol{y}').$$

Moreover, Φ is $L(\kappa_2 + 1)$-smooth, where $\kappa_2 = L/\mu_2$.

Proof. This is already proved in Lemma A.5 of Nouiehed et al. (2019). However, we present a bit different proof without using second-order Taylor expansion. To start, recall $\mathcal{Y}_{\boldsymbol{x}}^* := \arg\max_{\boldsymbol{y} \in \mathcal{Y}} g(\boldsymbol{x}; \boldsymbol{y})$. That is, we could choose any $\boldsymbol{y}^*(\boldsymbol{x}) \in \mathcal{Y}_{\boldsymbol{x}}^*$.

We first show the differentiability of Φ. Fix a unit vector $\boldsymbol{u} \in \mathcal{X} = \mathbb{R}^{d_x}$: $\|\boldsymbol{u}\| = 1$. Let any $h > 0$. We first claim that there exists a path $\boldsymbol{p} : (-h, h] \to \mathcal{Y} = \mathbb{R}^{d_y}$ which is continuous at $t = 0$ and $\boldsymbol{p}(t) \in \mathcal{Y}^*_{(\boldsymbol{x}+t\boldsymbol{u})}$. In fact, let $\boldsymbol{p}(t)$ be a projection of $\boldsymbol{y}^*(\boldsymbol{x})$ (that we chose) onto the set $\mathcal{Y}^*_{(\boldsymbol{x}+t\boldsymbol{u})}$. Then, $\boldsymbol{p}(0) = \boldsymbol{y}^*(\boldsymbol{x})$, and by Proposition 8, we have $\|\boldsymbol{p}(0) - \boldsymbol{p}(t)\| \leq \kappa_2 \|\boldsymbol{x} - (\boldsymbol{x} + t\boldsymbol{u})\| = \kappa_2 t$. This shows the continuity of $\boldsymbol{p}(t)$ at $t = 0$. Now, note that there exists a $t_1 \in (0, h)$ such that,

$$\begin{aligned}\Phi(\boldsymbol{x} + h\boldsymbol{u}) - \Phi(\boldsymbol{x}) &= g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(h)) - g(\boldsymbol{x}; \boldsymbol{p}(0)) \\ &= \left\{g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(h)) - g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(0))\right\} + \left\{g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(0)) - g(\boldsymbol{x}; \boldsymbol{p}(0))\right\} \\ &\geq 0 + \langle \nabla_1 g(\boldsymbol{x} + t_1\boldsymbol{u}; \boldsymbol{p}(0)), h\boldsymbol{u}\rangle ,\end{aligned}$$

by mean value theorem (applied to the first argument). We have the inequality at the last line because $g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(h)) \geq g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(0))$, since $\boldsymbol{p}(h) \in \mathcal{Y}^*_{(x+hu)}$. With a similar logic, there exists a $t_2 \in (0, h)$ such that,

$$\begin{aligned}\Phi(\boldsymbol{x} + h\boldsymbol{u}) - \Phi(\boldsymbol{x}) &= g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(h)) - g(\boldsymbol{x}; \boldsymbol{p}(0)) \\ &= \left\{g(\boldsymbol{x} + h\boldsymbol{u}; \boldsymbol{p}(h)) - g(\boldsymbol{x}; \boldsymbol{p}(h))\right\} + \left\{g(\boldsymbol{x}; \boldsymbol{p}(h)) - g(\boldsymbol{x}; \boldsymbol{p}(0))\right\} \\ &\leq \langle \nabla_1 g(\boldsymbol{x} + t_2\boldsymbol{u}; \boldsymbol{p}(h)), h\boldsymbol{u}\rangle + 0.\end{aligned}$$

To combine these two inequalities into a single line,

$$\langle \nabla_1 g(\boldsymbol{x} + t_1\boldsymbol{u}; \boldsymbol{p}(0)), \boldsymbol{u}\rangle \leq \frac{\Phi(\boldsymbol{x} + h\boldsymbol{u}) - \Phi(\boldsymbol{x})}{h} \leq \langle \nabla_1 g(\boldsymbol{x} + t_2\boldsymbol{u}; \boldsymbol{p}(h)), \boldsymbol{u}\rangle .$$

Using the continuity of $p(\cdot)$ and $\nabla_1 g(\cdot;\cdot)$ ($\because g$ has Lipschitz continuous gradient), we can deduce that the directional derivative of Φ in a direction $\boldsymbol{u}$ (denoted by $D_{\boldsymbol{u}}\Phi$) is in fact

$$D_{\boldsymbol{u}}\Phi(\boldsymbol{x}) = \langle \nabla_1 g(\boldsymbol{x}; \boldsymbol{y}^*(\boldsymbol{x})), \boldsymbol{u}\rangle ,$$

by taking the limit $h \to 0+$. Since $\boldsymbol{u}$ is arbitrary, we can conclude that $\nabla\Phi(\boldsymbol{x}) = \nabla_1 g(\boldsymbol{x}; \boldsymbol{y}^*(\boldsymbol{x}))$.

The proof of Lipschitz smoothness of Φ exactly follows the proof by Nouiehed et al. (2019). Consider any $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathcal{X}$. As in Proposition 8, choose any $\boldsymbol{y}_0^* \in \mathcal{Y}^*_{\boldsymbol{x}_0}$ and $\boldsymbol{y}_1^* \in \Pi_{\mathcal{Y}^*_{\boldsymbol{x}_1}}(\boldsymbol{y}_0^*)$. Then,

$$\begin{aligned}\|\nabla\Phi(\boldsymbol{x}_0) - \nabla\Phi(\boldsymbol{x}_1)\| &= \|\nabla_1 g(\boldsymbol{x}_0; \boldsymbol{y}_0^*) - \nabla_1 g(\boldsymbol{x}_1; \boldsymbol{y}_1^*)\| \\ &\leq \|\nabla_1 g(\boldsymbol{x}_0; \boldsymbol{y}_0^*) - \nabla_1 g(\boldsymbol{x}_1; \boldsymbol{y}_0^*)\| + \|\nabla_1 g(\boldsymbol{x}_1; \boldsymbol{y}_0^*) - \nabla_1 g(\boldsymbol{x}_1; \boldsymbol{y}_1^*)\| \\ &\leq L\left\{\|\boldsymbol{x}_0 - \boldsymbol{x}_1\| + \|\boldsymbol{y}_0^* - \boldsymbol{y}_1^*\|\right\} \\ &\leq L(1 + \kappa_2)\|\boldsymbol{x}_0 - \boldsymbol{x}_1\| .\end{aligned}$$

The last inequality holds because of Proposition 8. □

Proposition 10 ($\boldsymbol{x}$-side PŁ $\Rightarrow$ primal PŁ)**.** *Suppose $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is L-smooth and two-sided PŁ with constants μ_1 and μ_2. Then, g satisfies primal PŁ condition: the function $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}' \in \mathcal{Y}} g(\boldsymbol{x}; \boldsymbol{y}')$ is μ_1-PŁ. As a result, a smooth two-sided PŁ function is PŁ(Φ)-PŁ.*

Proof. See Lemma A.3 of Yang et al. (2020). □

Definition 2. *Consider $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Then, the point $(\boldsymbol{x}^*; \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}$ is called*

(i) a stationary point of g if $\nabla_1 g(\boldsymbol{x}^*; \boldsymbol{y}^*) = \nabla_2 g(\boldsymbol{x}^*; \boldsymbol{y}^*) = 0$.

(ii) a saddle point of g if $g(\boldsymbol{x}^*; \boldsymbol{y}) \leq g(\boldsymbol{x}^*; \boldsymbol{y}^*) \leq g(\boldsymbol{x}; \boldsymbol{y}^*)$ for all $\boldsymbol{x}, \boldsymbol{y}$.

(iii) a global minimax point of g if $g(\boldsymbol{x}^*; \boldsymbol{y}) \leq g(\boldsymbol{x}^*; \boldsymbol{y}^*) \leq \max_{\boldsymbol{y}'} g(\boldsymbol{x}; \boldsymbol{y}')$ for all $\boldsymbol{x}, \boldsymbol{y}$.

(iv) a global maximin point of g if $\min_{\boldsymbol{x}'} g(\boldsymbol{x}'; \boldsymbol{y}) \leq g(\boldsymbol{x}^*; \boldsymbol{y}^*) \leq g(\boldsymbol{x}; \boldsymbol{y}^*)$ for all $\boldsymbol{x}, \boldsymbol{y}$.

Proposition 11. *Consider a function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.*

(1) In general, a saddle point of g is a global minimax/maximin point.

(2) Let $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}} g(\boldsymbol{x};\boldsymbol{y})$ and $\Phi^* := \min_x \Phi(\boldsymbol{x})$ be well-defined. Let $\lambda > 0$ be a constant. In general, a point $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ is a global minimax point of g if and only if

$$V_\lambda(\boldsymbol{x}^*;\boldsymbol{y}^*) := \lambda[\Phi(\boldsymbol{x}) - \Phi^*] + [\Phi(\boldsymbol{x}) - g(\boldsymbol{x};\boldsymbol{y})] = 0.$$

(3) If g is smooth nonconvex-PŁ, then a global minimax point is a stationary point.

(4) If g is PŁ(Φ)-PŁ, then there exists a global minimax point $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ of g. As a result, if g is also smooth, then the point $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ is a stationary point.

(5) If g is smooth two-sided PŁ, every stationary point is a saddle point. As a result, there exists a saddle point $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ of g.

In particular, smooth two-sided PŁ functions enjoy the "minimax theorem," which establishes "minimax = maximin."

Proof. (1) (saddle point $\Rightarrow$ global minimax & global maximin) This is straightforward by the definitions: for any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\min_{\boldsymbol{x}'} g(\boldsymbol{x}';\boldsymbol{y}) \le g(\boldsymbol{x}^*;\boldsymbol{y}) \le g(\boldsymbol{x}^*;\boldsymbol{y}^*) \le g(\boldsymbol{x};\boldsymbol{y}^*) \le \max_{\boldsymbol{y}'} g(\boldsymbol{x};\boldsymbol{y}').$$

(2) (global minimax $\iff$ $V_\lambda = 0$) The terms $\Phi(\boldsymbol{x}) - \Phi^*$ and $\Phi(\boldsymbol{x}) - g(\boldsymbol{x};\boldsymbol{y})$ are non-negative. Hence, $V_\lambda(\boldsymbol{x};\boldsymbol{y})$ is non-negative, and $V_\lambda(\boldsymbol{x}^*;\boldsymbol{y}^*) = 0$ if and only if $\Phi^* = \Phi(\boldsymbol{x}^*) = g(\boldsymbol{x}^*;\boldsymbol{y}^*)$, which is equivalent to the global minimax point condition.

(3) (smooth nonconvex-PŁ: global minimax $\Rightarrow$ stationary) Suppose $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ is a global minimax point. Since $g(\boldsymbol{x}^*;\boldsymbol{y}) \le g(\boldsymbol{x}^*;\boldsymbol{y}^*)$ for any $\boldsymbol{y}$, $\Phi(\boldsymbol{x}^*) = \max_y g(\boldsymbol{x}^*;\boldsymbol{y}) = g(\boldsymbol{x}^*;\boldsymbol{y}^*)$. Thus, Φ has a minimum $g(\boldsymbol{x}^*;\boldsymbol{y}^*)$ at $\boldsymbol{x} = \boldsymbol{x}^*$. By Proposition 9, $\Phi(\cdot)$ is a differentiable function and we have

$$\nabla_1 g(\boldsymbol{x}^*;\boldsymbol{y}^*) = \nabla\Phi(\boldsymbol{x}^*) = 0.$$

Also, since a differentiable function $g(\boldsymbol{x}^*;\boldsymbol{y})$ has a maximum at $\boldsymbol{y} = \boldsymbol{y}^*$, we also have $\nabla_2 g(\boldsymbol{x}^*;\boldsymbol{y}^*) = 0$. Therefore, $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ is a stationary point.

(4) (PŁ(Φ)-PŁ: $\exists$ global minimax) Let $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}} \Phi(\boldsymbol{x})$ and $\boldsymbol{y}^* \in \arg\max_{\boldsymbol{y}} f(\boldsymbol{x}^*;\boldsymbol{y})$. Then, $f(\boldsymbol{x}^*,\boldsymbol{y}^*) = \Phi(\boldsymbol{x}^*) = \Phi^*$. as noted in (2), $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ is a global minimax point. By (3), it is in fact a stationary point, when g is smooth as well.

(5) (smooth two-sided PŁ: stationary $\Rightarrow$ saddle) Let $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ be a stationary point. By PŁ inequalities, for any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$0 = \|\nabla_2 g(\boldsymbol{x}^*;\boldsymbol{y}^*)\|^2 \ge 2\mu_2(\max_{\boldsymbol{y}} g(\boldsymbol{x}^*;\boldsymbol{y}) - g(\boldsymbol{x}^*;\boldsymbol{y}^*)) \ge 0,$$

$$0 = \|\nabla_1 g(\boldsymbol{x}^*;\boldsymbol{y}^*)\|^2 \ge 2\mu_1(g(\boldsymbol{x}^*;\boldsymbol{y}^*) - \min_{\boldsymbol{x}} g(\boldsymbol{x};\boldsymbol{y}^*)) \ge 0.$$

Since $\mu_1, \mu_2 > 0$, these imply $\max_{\boldsymbol{y}} g(\boldsymbol{x}^*;\boldsymbol{y}) = g(\boldsymbol{x}^*;\boldsymbol{y}^*) = \min_{\boldsymbol{x}} g(\boldsymbol{x};\boldsymbol{y}^*)$. Thus, $(\boldsymbol{x}^*;\boldsymbol{y}^*)$ is a saddle point. Note that (4) and Proposition 10 together proves the existence of a stationary point of g. Therefore, there must exists a saddle point, which is also pointed out by Guo et al. (2020, Lemma 8). This concludes the proof. □

We remark that, in the proof above, (3) is false for general (nonconvex-nonconcave) functions. Only *local* minimax point can ensure stationarity (Jin et al., 2020). As remarked by Jin et al. (2020) (Figure 2 of their paper), the function $xy - \cos(y)$ has non-stationary global minimax points $(0, \pm\pi)$.

The following two propositions are for showing that general two-sided PŁ function may not have a *differential Stackelberg equilibrium* defined as Li et al. (2022, Definition 3.1).

Proposition 12. *Let g be a μ-strongly convex function on $\mathbb{R}^n$. Consider any matrix $\boldsymbol{M} \in \mathbb{R}^{n\times m}$ with a positive rank. Suppose that θ is the smallest nonzero singular value of $\boldsymbol{M}$. Then $g(\boldsymbol{M}\boldsymbol{y})$ is a $\mu\theta^2$-PŁ function of $\boldsymbol{y} \in \mathbb{R}^m$.*

Proof. See Karimi et al. (2016, Appendix B) for the proof. □

Proposition 13. *Consider a twice continuously differentiable strongly-convex-strongly-concave function $h : \mathbb{R}^r \times \mathbb{R}^s \to \mathbb{R}$. That is, for some constants $\mu_1, \mu_2 > 0$, $h(\boldsymbol{x}; \boldsymbol{y})$ is μ_1-strongly-convex in $\boldsymbol{x}$ and $-h(\boldsymbol{x}; \boldsymbol{y})$ is μ_2-strongly-convex in $\boldsymbol{y}$. Let $(\boldsymbol{x}^*; \boldsymbol{y}^*)$ be the unique stationary point of h. Of course, it is a* ***differential Stackelberg equilibrium*** *of h. That is, if the hessian matrix $\nabla^2 h(\boldsymbol{x}^*; \boldsymbol{y}^*)$ at that point is written as*

$$\nabla^2 h(\boldsymbol{x}^*; \boldsymbol{y}^*) = \begin{bmatrix} \nabla^2_{1,1}\, h(\boldsymbol{x}^*; \boldsymbol{y}^*) & \nabla^2_{1,2}\, h(\boldsymbol{x}^*; \boldsymbol{y}^*) \\ \nabla^2_{2,1}\, h(\boldsymbol{x}^*; \boldsymbol{y}^*) & \nabla^2_{2,2}\, h(\boldsymbol{x}^*; \boldsymbol{y}^*) \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{B}^\top & -\boldsymbol{A} \end{bmatrix},$$

then $\boldsymbol{A}$ and $\boldsymbol{C} - \boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^\top$ are both positive definite matrices. Consider a function $g : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ defined by $g(\boldsymbol{x}; \boldsymbol{y}) = h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y})$ for some matrices $\boldsymbol{M} \in \mathbb{R}^{r \times p}$, $\boldsymbol{N} \in \mathbb{R}^{s \times q}$. Then, g is two-sided PŁ. Moreover, each stationary point of g may not be a differential Stackelberg equilibrium in general, for example, when $s < q$.

Proof. Because of Proposition 12, g is clearly a two-sided PŁ function.

If $(\boldsymbol{x}; \boldsymbol{y})$ is a stationary point of g, then it must be an element of an affine set $\{(\boldsymbol{x}; \boldsymbol{y}) \in \mathbb{R}^p \times \mathbb{R}^q : \boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^*; \boldsymbol{N}\boldsymbol{y} = \boldsymbol{y}^*\}$. This is because

$$\nabla g(\boldsymbol{x}; \boldsymbol{y}) = \begin{bmatrix} \nabla_1\, g(\boldsymbol{x}; \boldsymbol{y}) \\ \nabla_2\, g(\boldsymbol{x}; \boldsymbol{y}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}^\top \nabla_1\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) \\ \boldsymbol{N}^\top \nabla_2\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) \end{bmatrix} = \boldsymbol{0}$$

if and only if $\nabla_1\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) = \boldsymbol{0}$ and $\nabla_2\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) = \boldsymbol{0}$, being equivalent to $\boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^*$ and $\boldsymbol{N}\boldsymbol{y} = \boldsymbol{y}^*$. Furthermore, the hessian of g at $(\boldsymbol{x}; \boldsymbol{y})$ is

$$\begin{aligned} \nabla^2 g(\boldsymbol{x}; \boldsymbol{y}) &= \begin{bmatrix} \boldsymbol{M}^\top \nabla^2_{1,1}\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) \boldsymbol{M} & \boldsymbol{M}^\top \nabla^2_{1,2}\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) \boldsymbol{N} \\ \boldsymbol{N}^\top \nabla^2_{2,1}\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) \boldsymbol{M} & \boldsymbol{N}^\top \nabla^2_{2,2}\, h(\boldsymbol{M}\boldsymbol{x}; \boldsymbol{N}\boldsymbol{y}) \boldsymbol{N} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{M}^\top \boldsymbol{C} \boldsymbol{M} & \boldsymbol{M}^\top \boldsymbol{B} \boldsymbol{N} \\ (\boldsymbol{M}^\top \boldsymbol{B} \boldsymbol{N})^\top & -\boldsymbol{N}^\top \boldsymbol{A} \boldsymbol{N} \end{bmatrix}. \end{aligned}$$

If $s < q$, the $q \times q$ matrix $\boldsymbol{N}^\top \boldsymbol{A} \boldsymbol{N}$ cannot have a full rank, thereby it cannot be even invertible. This implies the stationary point $(\boldsymbol{x}; \boldsymbol{y})$ cannot be a differential Stackelberg equilibrium. □

B.2 Without-replacement sampling

In this subsection, we provide a useful proposition for the analysis of mini-batching approach under without-replacement sampling. We consider the case of mutually disjoint mini-batches in a whole epoch, not only applying without-replacement sampling to each individual mini-batch.

Consider a collection of n vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^d$. Suppose we randomly sample a permutation $\sigma : [n] \to [n]$; *i.e.*, $\sigma \sim \mathrm{Unif}(\mathbb{S}_n)$. Define

$$\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{v}_i \ \text{(sample mean)} \quad \text{and} \quad \tau^2 = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{v}_i - \boldsymbol{m}\|^2 \ \text{(sample variance)}.$$

Fix any $b \in [n]$ and let $n = b(q-1) + s$ for some integers $q \geq 1$ and $s \in [b]$. Now, divide the indices $[n]$ into q batches, with exactly b items per batch (except for the last batch, when $s < b$), as follows:

$$\begin{aligned} \mathcal{W}_1 &= \{\sigma(1), \cdots, \sigma(b)\}, \\ \mathcal{W}_2 &= \{\sigma(b+1), \cdots, \sigma(2b)\}, \\ &\vdots \\ \mathcal{W}_{q-1} &= \{\sigma(b(q-2)+1), \cdots, \sigma(b(q-1))\}, \\ \mathcal{W}_q &= \{\sigma(b(q-1)+1), \cdots, \sigma(n)\}. \end{aligned}$$

For each batch $\mathcal{W}_t$, define

$$\boldsymbol{w}_t = \frac{1}{|\mathcal{W}_t|} \sum_{i \in \mathcal{W}_t} \boldsymbol{v}_i \ \text{(batch mean)}.$$

For any $k \in [q-1]$, define

$$\boldsymbol{m}_k := \frac{1}{k}\sum_{t=1}^{k} \boldsymbol{w}_t \text{ (accumulative average of batch means over } 1 \le t \le k).$$

Of course, we may take $\boldsymbol{m}_q = \boldsymbol{m}$ (deterministically) for $k = q$. Then, due to the randomness of σ, we can obtain the mean (vector) and the variance (scalar) of $\boldsymbol{m}_k$ as follows.

Proposition 14 (Without-replacement sampling)**.** *Given the setup above, for any $k < q$ and $n > 1$,*

$$\mathbb{E}[\boldsymbol{m}_k] = \boldsymbol{m} \quad \textit{and} \quad \mathbb{E}\left[\|\boldsymbol{m}_k - \boldsymbol{m}\|^2\right] = \frac{(n-bk)}{bk(n-1)}\tau^2.$$

(Of course, if $k = q$ or $n = 1 = q$, $\mathbb{E}[\|\boldsymbol{m}_q - \boldsymbol{m}\|^2] = 0$ since $\boldsymbol{m}_q = \boldsymbol{m}$.)

Remark. As a special case, if $n = bq$ (namely, b divides n and $r = b$), then for any $k \le q$,

$$\mathbb{E}\left[\|\boldsymbol{m}_k - \boldsymbol{m}\|^2\right] = \frac{(q-k)}{k(n-1)}\tau^2.$$

If we further assume $b = s = 1$ and $q = n$, this proposition exactly recovers Lemma 1 of Mishchenko et al. (2020).

Proof of Proposition 14. Since σ is a uniformly randomly sampled permutation, it is easy to obtain that

$$\mathbb{E}[\boldsymbol{v}_{\sigma(i)}] = \mathbb{E}[\boldsymbol{w}_t] = \mathbb{E}[\boldsymbol{m}_k] = \boldsymbol{m},$$

for any $i \in [n]$, $t \in [q]$, and $k \in [q]$.

The covariances between $\boldsymbol{v}_{\sigma(i)}$'s can be deduced from the proof by Mishchenko et al. (2020, Lemma 1) as follows:

$$\mathrm{Cov}(\boldsymbol{v}_{\sigma(i)}, \boldsymbol{v}_{\sigma(j)}) := \mathbb{E}\left[\left\langle \boldsymbol{v}_{\sigma(i)} - \boldsymbol{m}, \boldsymbol{v}_{\sigma(j)} - \boldsymbol{m}\right\rangle\right] = \begin{cases} -\frac{\tau^2}{n-1}, & \text{if } i \neq j, \\ \tau^2 & \text{if } i = j. \end{cases}$$

Thus, for each $t \in [q]$, the variance of $\boldsymbol{w}_t$ is obtained as

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{w}_t - \boldsymbol{m}\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{|\mathcal{W}_t|}\sum_{i\in\mathcal{W}_t}(\boldsymbol{v}_i - \boldsymbol{m})\right\|^2\right] \\
&= \frac{1}{|\mathcal{W}_t|^2}\left\{\sum_{i\in\mathcal{W}_t}\mathbb{E}\left[\|\boldsymbol{v}_i - \boldsymbol{m}\|^2\right] + \sum_{\substack{i,j\in\mathcal{W}_t \\ i\neq j}}\mathrm{Cov}(\boldsymbol{v}_i, \boldsymbol{v}_j)\right\} \\
&= \frac{1}{|\mathcal{W}_t|^2}\left\{|\mathcal{W}_t|\tau^2 + |\mathcal{W}_t|(|\mathcal{W}_t|-1)\left(-\frac{\tau^2}{n-1}\right)\right\} = \frac{n-|\mathcal{W}_t|}{|\mathcal{W}_t|(n-1)}\tau^2,
\end{aligned}$$

which can also be directly deduced by Lemma 1 of Mishchenko et al. (2020). We notice that this does not depends on the size of the batch $\mathcal{W}_t$.

Next, we look at the covariances between distinct $\boldsymbol{w}_t$'s. For a pair of distinct integers $t, u \in [q]$, by the bi-linearity of covariance,

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{w}_t, \boldsymbol{w}_u) &= \frac{1}{|\mathcal{W}_t|\cdot|\mathcal{W}_u|}\sum_{(i,j)\in\mathcal{W}_t\times\mathcal{W}_u}\mathrm{Cov}(\boldsymbol{v}_i, \boldsymbol{v}_j) \\
&= \frac{1}{|\mathcal{W}_t|\cdot|\mathcal{W}_u|}\sum_{(i,j)\in\mathcal{W}_t\times\mathcal{W}_u}\left(-\frac{\tau^2}{n-1}\right) = -\frac{\tau^2}{n-1}.
\end{aligned}$$

The second equality holds because $\mathcal{W}_t$ and $\mathcal{W}_u$ are a disjoint set of integers whenever $t \neq u$.

Now, fix any $k \in [q-1]$. Note that, by our mini-batching strategy, $|\mathcal{W}_t| = b$ for every $t < q$. Therefore, by definition of $\boldsymbol{m}_k$,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{m}_k - \boldsymbol{m}\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^{k}(\boldsymbol{w}_t - \boldsymbol{m})\right\|^2\right] \\
&= \frac{1}{k^2}\left\{\sum_{t=1}^{k}\mathbb{E}\left[\|\boldsymbol{w}_t - \boldsymbol{m}\|^2\right] + \sum_{\substack{t,u\in[k]\\ t\neq u}}\mathrm{Cov}(\boldsymbol{w}_t, \boldsymbol{w}_u)\right\} \\
&= \frac{1}{k^2}\left\{k\cdot\left(\frac{n-b}{b(n-1)}\tau^2\right) + k(k-1)\cdot\left(-\frac{\tau^2}{n-1}\right)\right\} = \frac{n-bk}{bk(n-1)}\tau^2.
\end{aligned}$$

□

B.3 Basic recurrence inequality

In this subsection, we present a basic result of a recurrence inequality. It serves as a stepping-stone of our convergence bound, particularly at the end of the proof (Appendix C.5).

Proposition 15. *Let $\{a_k\}_{k=1}^{\infty}$ be a sequence of non-negative numbers satisfying the following recurrence inequality:*

$$a_{k+1} \leq (1-b\eta)a_k + c\eta^{m+1},$$

where b, c, and η are non-negative real numbers such that $b\eta \in (0,1)$, and m is a non-negative integer. Then, for any integer $K \geq 1$, we have

$$a_{K+1} \leq (1-b\eta)^K a_1 + c\eta^m/b.$$

Proof. We proceed with induction on $K = 0, 1, 2, \cdots$. Note that

$$a_1 \leq (1-b\eta)^0 a_1 + c\eta^m/b.$$

This shows the case when $K = 0$. On the other hand, if $K \geq 1$, by an inductive assumption,

$$\begin{aligned}
a_{K+1} &\leq (1-b\eta)a_K + c\eta^{m+1} \\
&\leq (1-b\eta)\cdot\left((1-b\eta)^{K-1}a_1 + c\eta^m/b\right) + c\eta^{m+1} \\
&= (1-b\eta)^K a_1 + c\eta^m/b.
\end{aligned}$$

□

C PROOFS FOR (MINI-BATCH) SIMULTANEOUS SGDA WITH RANDOM RESHUFFLING

In this appendix, we provide a convergence analysis for the mini-batch **sim**SGDA-RR (Algorithm 2) on both general nonconvex-PŁ problems and PŁ(Φ)-PŁ problems. The two cases mostly share the same proof strategies; they only diverge at the end of the proofs. The proof is long; we first provide the sketch of proof in subsection C.1; then, we provide the full proof by dividing it into 4 follow-up subsections of this appendix.

The proof for the alternating counterpart (altSGDA-RR) can be done with some modifications illustrated in Appendix D.

C.1 WARM-UP: PROOF SKETCH FOR $b = 1$

Here we simply consider the proofs of Theorem 1 and 2 for **sim**SGDA-RR, which is a fully stochastic case (mini-batches of size $b = 1$). The proofs for altSGDA-RR can be done with slight modifications.

We start the proof by aggregating all updates throughout an epoch to obtain an "epoch-wise" update:

$$\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_0^k - n\alpha \boldsymbol{g}^k, \quad \boldsymbol{g}^k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla_1 f_{\sigma_k(i)}(\boldsymbol{z}_{i-1}^k),$$
$$\boldsymbol{y}_0^{k+1} = \boldsymbol{y}_0^k + n\beta \boldsymbol{h}^k, \quad \boldsymbol{h}^k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla_2 f_{\sigma_k(i)}(\boldsymbol{z}_{i-1}^k).$$

The reason is that the sampled components in each epoch are dependent to each other so that it is much harder to deal with each iteration individually. The strategy of update-aggregation is quite general for analysis of optimization algorithms involving without-replacement sampling (Ahn et al., 2020; Mishchenko et al., 2020; Nguyen et al., 2021; Das et al., 2022). We assume that the intermediate iterates $\boldsymbol{z}_1^k, \dots, \boldsymbol{z}_n^k$ stay close to the starting iterate $\boldsymbol{z}_0^k$ of an epoch k, which can be ensured by small step sizes. Then, we can approximate the aggregated epoch of SGDA-RR as a step of simGDA applied to $f = \frac{1}{n}\sum_{i=1}^n f_i$, with approximations of $\boldsymbol{g}^k \approx \nabla_1 f(\boldsymbol{z}_0^k)$ and $\boldsymbol{h}^k \approx \nabla_2 f(\boldsymbol{z}_0^k)$.

With Assumptions 1 and 4, note that the primal function $\Phi(\cdot)$ is $(L+L^2/\mu_2)$-smooth (Proposition 9). Applying this and L-smoothness of $-f$, we can have the following inequality (Lemma 16):

$$\begin{aligned} V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k) \le & -((\lambda+1)/2)\, n\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (\lambda+1)n\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\ & + (n\alpha/2)\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 - (n\beta/2)\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \\ & + (\lambda+1/2)\, n\alpha \left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + (n\beta/2)\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2. \end{aligned}$$

Hence, to guarantee the fast decrease of $V_\lambda(\boldsymbol{z}_0^k)$, it is important to control the "noise" terms for GDA approximations, $\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2$ and $\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2$, in the last line of inequality above. By applying the tools for without-replacement sampling (Proposition 14), we can actually upper-bound the conditional expectations of both noise terms by

$$2L^2 n(n+A)\left(\alpha^2\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \beta^2\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2\right) + 2L^2 n(\alpha^2+\beta^2)B. \quad \text{(Lemma 17 \& 18)}$$

Then, by taking advantage of several properties of smooth nonconvex-PŁ functions (*e.g.*, Propositions 7, 8, and 9) and some small-step-size assumptions (*e.g.*, $\beta = \mathcal{O}(1/nL)$, $\beta/\alpha = r \gtrsim \kappa_2^2$), we eventually have

$$\mathbb{E}\left[V_\lambda(\boldsymbol{z}_0^{k+1})\right] - \mathbb{E}\left[V_\lambda(\boldsymbol{z}_0^k)\right] \le -n\alpha\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] - (L\kappa_2 n\alpha/2)\mathbb{E}\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right] + C\alpha^3,$$

where $C \ge 0$ is a constant (with respect to k) depending on L, n, B, and $r = \beta/\alpha$. (Lemma 20). We note that the step size ratio $r \gtrsim \kappa_2^2$ is crucial for showing that the coefficient in front of the term $\mathbb{E}\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right]$ is non-positive: even if it is possible with $r \lesssim \kappa_2^2$, we must assume that κ_2 upper-bounded by a positive numerical constant, which is not desirable for showing convergence bounds. Thus, we expect that a different proof strategy should be applied to avoid the requirement $r \gtrsim \kappa_2^2$ on the step size ratio.

The proofs of Theorems 1 and 2 diverge from here. The rest of the proof is mostly about choosing appropriate step sizes and solving the recurrence inequalities.

The full proof of Theorems 4 and 5 starts from the following subsection.

C.2 EPOCH-WISE REPRESENTATIONS AND BOUNDING NOISE TERMS

Before starting the proof, we again remark that we assume that the mini-batch size b divides the number of components n (namely, $q := n/b$ is a positive integer) for simplicity: thus, readers who just want to read proofs for fully stochastic case (*i.e.*, $b = 1$) can substitute n to every q. Also, there is no problem in treating any fraction with a positive numerator and a zero denominator as $+\infty$. Moreover, we simply regard $(q-1)/(n-1) = 1$ when $n = 1$ ($\because q = 1 = n$ in this case).

We start the proof by aggregating all updates throughout an epoch to obtain an "epoch-wise" update equation. The reason is that the sampled components in each epoch are dependent to each other, so that it is much harder to deal with each iteration individually. At iteration $t \in [n/b] = [q]$ of epoch $k \in [K]$, we use a mini-batch

$$\mathcal{B}_t^k := \{\sigma_k(b(t-1)+1), \dots, \sigma_k(bt)\}.$$

Denote the averaged (partial) gradients evaluated at a point $\boldsymbol{z} = (\boldsymbol{x}; \boldsymbol{y})$ over the mini-batch as

$$\boldsymbol{g}_t^k(\boldsymbol{z}) := \frac{1}{b} \sum_{i \in \mathcal{B}_t^k} \nabla_1 f_i(\boldsymbol{z}), \quad \boldsymbol{h}_t^k(\boldsymbol{z}) := \frac{1}{b} \sum_{i \in \mathcal{B}_t^k} \nabla_2 f_i(\boldsymbol{z}).$$

By Assumption 1, $\boldsymbol{g}_t^k$ and $\boldsymbol{h}_t^k$ are L-Lipschitz continuous. Computing the average of them over a whole epoch $(\boldsymbol{z}_0^k, \cdots, \boldsymbol{z}_{q-1}^k)$, we define

$$\boldsymbol{g}^k := \frac{1}{q} \sum_{t=1}^{q} \boldsymbol{g}_t^k(\boldsymbol{z}_{t-1}^k), \quad \boldsymbol{h}^k := \frac{1}{q} \sum_{t=1}^{q} \boldsymbol{h}_t^k(\boldsymbol{z}_{t-1}^k).$$

Then, by summing up the updates in the epoch k, we can summarize the epoch as follows.

$$\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_0^k - q\alpha \boldsymbol{g}^k, \quad \boldsymbol{y}_0^{k+1} = \boldsymbol{y}_0^k + q\beta \boldsymbol{h}^k. \tag{simSGDA-RR}$$

We may assume that the intermediate iterates $\boldsymbol{z}_1^k, \dots, \boldsymbol{z}_q^k$ stay close to the starting iterate $\boldsymbol{z}_0^k$ of an epoch k, which results from, *e.g.*, small step sizes. Then, we can approximate the aggregated epoch of SGDA-RR as a step of simGDA applied to $f = \frac{1}{n}\sum_{i=1}^n f_i$: $\boldsymbol{g}^k \approx \nabla_1 f(\boldsymbol{z}_0^k), \quad \boldsymbol{h}^k \approx \nabla_2 f(\boldsymbol{z}_0^k)$. In other words,

$$\boldsymbol{x}_0^{k+1} \approx \boldsymbol{x}_0^k - q\alpha \nabla_1 f(\boldsymbol{z}_0^k), \quad \boldsymbol{y}_0^{k+1} \approx \boldsymbol{y}_0^k + q\beta \nabla_2 f(\boldsymbol{z}_0^k), \tag{$\approx$simGDA}$$

With Assumptions 1, 3 and 4, we can yield a naive (but complicated) upper bound of the gap $V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k)$, only applying the smoothness of Φ and $-f$, without any assumptions on step sizes.

Lemma 16. *Suppose that Assumptions 1, 3 and 4 hold. Let $\kappa_2 = L/\mu_2$, where μ_2 is PŁ constant of $-f(\boldsymbol{x}; \cdot)$. Then, the mini-batch simSGDA-RR satisfies that*

$$\begin{aligned} V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k) \le & -\left(\frac{\lambda+1}{2}\right) q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (\lambda+1) q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\ & + \frac{q\alpha}{2} \left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 - \frac{q\beta}{2} \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \\ & + \left(\lambda + \frac{1}{2}\right) q\alpha \left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{q\beta}{2} \left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \\ & - \left[\lambda - \{(\lambda+1)(\kappa_2+1)+1\} Lq\alpha\right] \frac{q\alpha}{2} \left\|\boldsymbol{g}^k\right\|^2 - (1 - Lq\beta)\frac{q\beta}{2} \left\|\boldsymbol{h}^k\right\|^2. \end{aligned} \tag{7}$$

Proof. By definition of V_λ, the following equation holds:

$$V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k) = (\lambda+1)\left[\Phi(\boldsymbol{x}_0^{k+1}) - \Phi(\boldsymbol{x}_0^k)\right] + \left[f(\boldsymbol{z}_0^k) - f(\boldsymbol{z}_0^{k+1})\right]. \tag{8}$$

First, we seek for an upper bound of $\Phi(\boldsymbol{x}_0^{k+1}) - \Phi(\boldsymbol{x}_0^k)$. By Proposition 9, Φ is $L(\kappa_2+1)$-smooth. Hence, we have

$$\begin{aligned}
\Phi(\boldsymbol{x}_0^{k+1}) - \Phi(\boldsymbol{x}_0^k) &\leq \left\langle \nabla\Phi(\boldsymbol{x}_0^k), \boldsymbol{x}_0^{k+1} - \boldsymbol{x}_0^k \right\rangle + \frac{L(\kappa_2+1)}{2}\left\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}_0^k\right\|^2 \\
&= -q\alpha\left\langle \nabla\Phi(\boldsymbol{x}_0^k), \boldsymbol{g}^k \right\rangle + \frac{L(\kappa_2+1)}{2} q^2\alpha^2 \left\|\boldsymbol{g}^k\right\|^2 \\
&= -\frac{q\alpha}{2}\left\{\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + \left\|\boldsymbol{g}^k\right\|^2 - \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \boldsymbol{g}^k\right\|^2\right\} + \frac{L(\kappa_2+1)}{2} q^2\alpha^2 \left\|\boldsymbol{g}^k\right\|^2 \\
&= -\frac{q\alpha}{2}\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + \frac{q\alpha}{2}\left\|\nabla\Phi(\boldsymbol{x}_0^k) - \boldsymbol{g}^k\right\|^2 - \frac{q\alpha}{2}(1 - L(\kappa_2+1)q\alpha)\left\|\boldsymbol{g}^k\right\|^2 \\
&\leq -\frac{q\alpha}{2}\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + q\alpha\left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + q\alpha\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\
&\quad - \frac{q\alpha}{2}(1 - L(\kappa_2+1)q\alpha)\left\|\boldsymbol{g}^k\right\|^2. \qquad (9)
\end{aligned}$$

The third line is due to polarization equality[8] and the last inequality applies Young's inequality.[9]

Next, applying Assumption 1, L-smoothness of $-f(\cdot;\cdot)$ yields an upper bound of $f(\boldsymbol{z}_0^k) - f(\boldsymbol{z}_0^{k+1})$.

$$\begin{aligned}
f(\boldsymbol{z}_0^k) - f(\boldsymbol{z}_0^{k+1}) &\leq -\left\langle \nabla f(\boldsymbol{z}_0^k), \boldsymbol{z}_0^{k+1} - \boldsymbol{z}_0^k \right\rangle + \frac{L}{2}\left\|\boldsymbol{z}_0^{k+1} - \boldsymbol{z}_0^k\right\|^2 \\
&= -\left\langle \nabla_1 f(\boldsymbol{z}_0^k), \boldsymbol{x}_0^{k+1} - \boldsymbol{x}_0^k \right\rangle - \left\langle \nabla_2 f(\boldsymbol{z}_0^k), \boldsymbol{y}_0^{k+1} - \boldsymbol{y}_0^k \right\rangle \\
&\quad + \frac{L}{2}\left\|\boldsymbol{x}_0^{k+1} - \boldsymbol{x}_0^k\right\|^2 + \frac{L}{2}\left\|\boldsymbol{y}_0^{k+1} - \boldsymbol{y}_0^k\right\|^2 \\
&= q\alpha\left\langle \nabla_1 f(\boldsymbol{z}_0^k), \boldsymbol{g}^k \right\rangle - q\beta\left\langle \nabla_2 f(\boldsymbol{z}_0^k), \boldsymbol{h}^k \right\rangle + \frac{L}{2}q^2\alpha^2\left\|\boldsymbol{g}^k\right\|^2 + \frac{L}{2}q^2\beta^2\left\|\boldsymbol{h}^k\right\|^2 \\
&= \frac{q\alpha}{2}\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 - \frac{q\alpha}{2}\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{q\alpha}{2}(1 + Lq\alpha)\left\|\boldsymbol{g}^k\right\|^2 \\
&\quad - \frac{q\beta}{2}\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{q\beta}{2}\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 - \frac{q\beta}{2}(1 - Lq\beta)\left\|\boldsymbol{h}^k\right\|^2. \qquad (10)
\end{aligned}$$

The last equality is due to polarization equality. Lastly, substituting (9) and (10) to (8) finishes the proof. □

We remark that, the last two terms of the inequality (7) can be simply ignored by applying small enough step sizes. However, the terms in the third line of (7) are non-negatives terms related to the "noise" of approximation $\boldsymbol{g}^k \approx \nabla_1 f(\boldsymbol{z}_0^k)$, $\boldsymbol{h}^k \approx \nabla_2 f(\boldsymbol{z}_0^k)$. Hence, it is important to control the noise terms $\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2$ and $\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2$ to guarantee a fast decrease of $V_\lambda(\boldsymbol{z}_0^k)$.

Lemma 17. *For mini-batch simSGDA-RR, define*

$$G_k := \frac{1}{q}\sum_{t=1}^{q}\left\|\boldsymbol{z}_{t-1}^k - \boldsymbol{z}_0^k\right\|^2. \qquad (11)$$

With Assumption 1, then

$$\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \leq L^2 G_k \quad \textit{and} \quad \left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \leq L^2 G_k.$$

As a side remark, $G_k = 0$ when $q = 1$ and, in particular, $n = 1$.

[8] For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - \|\boldsymbol{a} - \boldsymbol{b}\|^2$.

[9] For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, $\|\boldsymbol{a} + \boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$.

Proof. Recall that $\frac{1}{q}\sum_{t=1}^{q} \boldsymbol{g}_t^k(\boldsymbol{z}) = \nabla_1 f(\boldsymbol{z})$ and $\frac{1}{q}\sum_{t=1}^{q} \boldsymbol{h}_t^k(\boldsymbol{z}) = \nabla_2 f(\boldsymbol{z})$. By Lipschitz continuity and Jensen's inequality,[10]

$$\begin{aligned}\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 &= \left\|\frac{1}{q}\sum_{t=1}^{q}\left[\boldsymbol{g}_t^k(\boldsymbol{z}_{t-1}^k) - \boldsymbol{g}_t^k(\boldsymbol{z}_0^k)\right]\right\|^2 \\ &\leq \frac{1}{q}\sum_{t=1}^{q}\left\|\boldsymbol{g}_t^k(\boldsymbol{z}_{t-1}^k) - \boldsymbol{g}_t^k(\boldsymbol{z}_0^k)\right\|^2 \leq \frac{L^2}{q}\sum_{t=1}^{q}\left\|\boldsymbol{z}_{t-1}^k - \boldsymbol{z}_0^k\right\|^2.\end{aligned}$$

Similarly,

$$\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \leq \frac{L^2}{q}\sum_{t=1}^{q}\left\|\boldsymbol{z}_{t-1}^k - \boldsymbol{z}_0^k\right\|^2.$$

This concludes the proof. □

Thanks to the lemma, it suffices to bound the term G_k. One can notice that it also represents how far the intermediate iterates $\boldsymbol{z}_t^k$ are from the pivot $\boldsymbol{z}_0^k$ in average. Before moving on, we define an algorithm-specific symbol denoting a conditional expectation.

Definition 3. *We denote a conditional expectation of a random variable X given all iterates of the first $k-1$ epochs by $\mathbb{E}_k[X] = \mathbb{E}[X|\boldsymbol{z}_0^1, \boldsymbol{z}_1^1, \ldots, \boldsymbol{z}_n^{k-1}]$. In particular, if $k=1$, it boils down to a conditional expectation given only the initial iterate $\boldsymbol{z}_0^1$.*

We get an upper bound of a (conditional) expectation $\mathbb{E}_k[G_k]$ in the following lemma, which extends a lemma of Nguyen et al. (2021, Lemma 6) to our minimax problems.

Lemma 18. *Suppose that Assumptions 1 and 2 hold. Assume that the permutation σ_k is sampled uniformly at random from $\mathbb{S}_n$. Then, for any step sizes α, β satisfying $\alpha^2 + \beta^2 \leq \frac{1}{3q(q-1)L^2}$, the iterates $\{\boldsymbol{z}_t^k\}_{t=0}^{q-1}$ of the k-th epoch of mini-batch simSGDA-RR satisfies (for $n > 1$)*

$$\mathbb{E}_k G_k \leq 2\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left(\alpha^2\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \beta^2\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2\right) + \frac{2q(q-1)}{n-1}(\alpha^2+\beta^2)B.$$

Proof. Note that $G_k = 0$ when $q = 1$ by its definition. From now, we may assume $q > 1$ and $n > 1$ in this proof. By summing the first $t \in [q-1]$ updates of the k-th epoch of mini-batch simSGDA-RR, we have

$$\boldsymbol{x}_t^k = \boldsymbol{x}_0^k - t\alpha\left(\frac{1}{t}\sum_{j=1}^{t}\boldsymbol{g}_j^k(\boldsymbol{z}_{j-1}^k)\right), \quad \boldsymbol{y}_t^k = \boldsymbol{y}_0^k + t\beta\left(\frac{1}{t}\sum_{j=1}^{t}\boldsymbol{h}_j^k(\boldsymbol{z}_{j-1}^k)\right).$$

Then we can bound the following squared distance.

$$\begin{aligned}&\left\|\boldsymbol{x}_t^k - \boldsymbol{x}_0^k\right\|^2 = \alpha^2 t^2\left\|\frac{1}{t}\sum_{j=1}^{t}\boldsymbol{g}_j^k(\boldsymbol{z}_{j-1}^k)\right\|^2 \\ &\leq 3\alpha^2 t^2\left[\left\|\frac{1}{t}\sum_{j=1}^{t}\left[\boldsymbol{g}_j^k(\boldsymbol{z}_{j-1}^k) - \boldsymbol{g}_j^k(\boldsymbol{z}_0^k)\right]\right\|^2 + \left\|\frac{1}{t}\sum_{j=1}^{t}\boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2\right] \\ &\leq 3\alpha^2 t\sum_{j=1}^{t}\left\|\boldsymbol{g}_j^k(\boldsymbol{z}_{j-1}^k) - \boldsymbol{g}_j^k(\boldsymbol{z}_0^k)\right\|^2 + 3\alpha^2 t^2\left[\left\|\frac{1}{t}\sum_{j=1}^{t}\boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2\right] \\ &\leq 3\alpha^2 L^2 t\cdot\sum_{j=1}^{t}\left\|\boldsymbol{z}_{j-1}^k - \boldsymbol{z}_0^k\right\|^2 + 3\alpha^2 t^2\left[\left\|\frac{1}{t}\sum_{j=1}^{t}\boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2\right]\end{aligned}$$

[10] For any n vectors $a_1, \cdots, a_n$, $\left\|\frac{1}{n}\sum_{j=1}^{n} a_j\right\|^2 \leq \frac{1}{n}\sum_{j=1}^{n}\|a_j\|^2$.

$$\leq 3\alpha^2 L^2 t \cdot qG_k + 3\alpha^2 t^2 \left[\left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + \left\| \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 \right]. \tag{12}$$

The second and third lines are due to Jensen's inequality. The fourth line is due to L-Lipschitz continuity of $\boldsymbol{g}_j^k$. Likewise,

$$\left\| \boldsymbol{y}_t^k - \boldsymbol{y}_0^k \right\|^2 \leq 3\beta^2 L^2 t \cdot qG_k + 3\beta^2 t^2 \left[\left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\| \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 \right]. \tag{13}$$

Summing up (12) and (13),

$$\begin{aligned} &\left\| \boldsymbol{z}_t^k - \boldsymbol{z}_0^k \right\|^2 = \left\| \boldsymbol{x}_t^k - \boldsymbol{x}_0^k \right\|^2 + \left\| \boldsymbol{y}_t^k - \boldsymbol{y}_0^k \right\|^2 \\ &\leq 3(\alpha^2 + \beta^2) L^2 tqG_k + 3\alpha^2 t^2 \left[\left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + \left\| \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 \right] \\ &\quad + 3\beta^2 t^2 \left[\left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\| \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 \right]. \end{aligned} \tag{14}$$

Taking (conditional) expectation $\mathbb{E}_k$ (given $\boldsymbol{z}_0^k$) to inequality (14),

$$\begin{aligned} &\mathbb{E}_k \left\| \boldsymbol{z}_t^k - \boldsymbol{z}_0^k \right\|^2 \\ &\overset{(14)}{\leq} 3(\alpha^2 + \beta^2) L^2 tq \cdot (\mathbb{E}_k[G_k]) + 3\alpha^2 t^2 \left\| \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + 3\beta^2 t^2 \left\| \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 \\ &\quad + 3\alpha^2 t^2 \mathbb{E}_k \left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + 3\beta^2 t^2 \mathbb{E}_k \left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2. \end{aligned} \tag{15}$$

Here we take advantage of the without-replacement sampling. Putting $\nabla_s f_i(\boldsymbol{z}_0^k) \mapsto \boldsymbol{v}_i$ ($s \in \{1, 2\}$), one can realize a correspondence between the quantities that arise from our algorithm and the symbols in Appendix B.2: for $s = 1$ ($\nabla_1 f_i(\boldsymbol{z}_0^k) \mapsto \boldsymbol{v}_i$),

$$\boldsymbol{m} = \nabla_1 f(\boldsymbol{z}_0^k), \quad \tau^2 \leq A \left\| \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + B, \quad \boldsymbol{w}_t = \boldsymbol{g}_t^k(\boldsymbol{z}_0^k), \quad \boldsymbol{m}_t = \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{g}_j^k(\boldsymbol{z}_0^k),$$

and for $s = 2$ ($\nabla_2 f_i(\boldsymbol{z}_0^k) \mapsto \boldsymbol{v}_i$),

$$\boldsymbol{m} = \nabla_2 f(\boldsymbol{z}_0^k), \quad \tau^2 \leq A \left\| \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + B, \quad \boldsymbol{w}_t = \boldsymbol{h}_t^k(\boldsymbol{z}_0^k), \quad \boldsymbol{m}_t = \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k).$$

The upper bounds of τ^2's come from Assumption 2. Then by Proposition 14, for any $t \leq q$,

$$t^2 \mathbb{E}_k \left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{g}_j^k(\boldsymbol{z}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 \leq \frac{t(q-t)}{n-1} \left(A \left\| \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + B \right),$$

$$t^2 \mathbb{E}_k \left\| \frac{1}{t} \sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 \leq \frac{t(q-t)}{n-1} \left(A \left\| \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + B \right).$$

Putting these to the inequality (15),

$$\begin{aligned} \mathbb{E}_k \left\| \boldsymbol{z}_t^k - \boldsymbol{z}_0^k \right\|^2 &\leq 3(\alpha^2 + \beta^2) \left[L^2 tq \mathbb{E}_k[G_k] + \frac{t(q-t)}{n-1} B \right] \\ &\quad + 3 \left(\alpha^2 \left\| \nabla_1 f(\boldsymbol{z}_0^k) \right\|^2 + \beta^2 \left\| \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 \right) \left[t^2 + \frac{t(q-t)}{n-1} A \right]. \end{aligned}$$

Taking an average of the inequality above over $0 \le t \le q-1$,

$$\begin{aligned}
\mathbb{E}_k G_k &= \frac{1}{q}\sum_{t=0}^{q-1} \mathbb{E}_k \left\| \boldsymbol{z}_t^k - \boldsymbol{z}_0^k \right\|^2 \\
&\le \frac{3q(q-1)}{2}(\alpha^2+\beta^2)L^2 \mathbb{E}_k G_k + (\alpha^2+\beta^2)\frac{q^2-1}{2(n-1)}B \\
&\quad + \left(\alpha^2 \left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \beta^2 \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2\right)\left(\frac{(q-1)(2q-1)}{2} + \frac{q^2-1}{2(n-1)}A\right), \qquad (16)
\end{aligned}$$

where we used the facts $\sum_{t=0}^{q-1} t = \frac{q(q-1)}{2}$, $\frac{1}{q}\sum_{t=0}^{q-1} t^2 = \frac{(q-1)(2q-1)}{6}$, and $\frac{1}{q}\sum_{t=0}^{q-1}\frac{t(q-t)}{n-1} = \frac{q^2-1}{6(n-1)}$. Since we assumed that $\alpha^2+\beta^2 \le \frac{1}{3q(q-1)L^2}$, we have $1 \le 2\left(1 - \frac{3q(q-1)L^2}{2}(\alpha^2+\beta^2)\right)$. Using this,

$$\begin{aligned}
\mathbb{E}_k G_k &\le 2\left(1 - \frac{3q(q-1)L^2}{2}(\alpha^2+\beta^2)\right)\mathbb{E}_k G_k \\
&\overset{(16)}{\le} \left((q-1)(2q-1) + \frac{q^2-1}{(n-1)}A\right)\left(\alpha^2\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \beta^2\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2\right) + \frac{q^2-1}{n-1}(\alpha^2+\beta^2)B \\
&\le 2\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left(\alpha^2\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \beta^2\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2\right) + \frac{2q(q-1)}{n-1}(\alpha^2+\beta^2)B,
\end{aligned}$$

where the last inequality used $(q-1)(2q-1) \le 2q^2$ and $q+1 \le 2q$ for $q \ge 1$. □

C.3 Recurrence inequalities for general smooth nonconvex-PŁ objective

Subsequently, we obtain recurrence inequalities about (expected) potential function $\mathbb{E}_k[V_\lambda(\boldsymbol{z}_0^k)]$ for nonconvex-PŁ problem. Since PŁ(Φ)-PŁ problem is a subclass of nonconvex-PŁ problem, the recurrence relations can serve as stepping-stones of our convergence rates.

We introduce some assumptions on *small* step sizes which enable us to get rid of a few troublesome terms from our bound. On top of that, combining the PŁ condition (Assumption 4) with Lemmas 16, 17, and 18, we eventually obtain a much more concise bound on the expected per-epoch change of V_λ. This simple recurrence inequality becomes the key to proving our convergence bounds.

Lemma 19. *Suppose that Assumptions 1, 2, 3, and 4 hold. Assume that the step sizes α and β satisfy*

$$\alpha \le \frac{\lambda}{\{(\lambda+1)(\kappa_2+1)+1\}qL}, \quad \beta \le \frac{1}{qL}, \quad \alpha^2+\beta^2 \le \frac{1}{3q(q-1)L^2}, \qquad (17)$$

and the condition

$$C_0 := q\beta - 2L^2 q\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left((2\lambda+1)\alpha+\beta\right)\beta^2 \ge 0$$

as well. Then, the iterates of mini-batch simSGDA-RR satisfy

$$\mathbb{E}_k[V_\lambda(\boldsymbol{z}_0^{k+1})] - V_\lambda(\boldsymbol{z}_0^k) \le -C_1\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 - C_2\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right] + C_3$$

where

$$\begin{aligned}
C_1 &= \left(\frac{\lambda-1}{2}\right)q\alpha - 2L^2 q\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left((2\lambda+1)\alpha+\beta\right)\alpha^2, \\
C_2 &= \mu_2 C_0 - 2(\lambda+2)L\kappa_2 q\alpha - 4L^3\kappa_2 q\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left((2\lambda+1)\alpha+\beta\right)\alpha^2 \\
&= \mu_2 q\beta - 2(\lambda+2)L\kappa_2 q\alpha - 2L^2\mu_2 q\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left((2\lambda+1)\alpha+\beta\right)\left(2\kappa_2^2\alpha^2+\beta^2\right), \\
C_3 &= \left(\frac{L^2q^2(q-1)}{n-1}\right)\left((2\lambda+1)\alpha+\beta\right)(\alpha^2+\beta^2)B.
\end{aligned}$$

Proof. The first two inequalities of (17) eliminate the last two terms on the right-hand side of the inequality in Lemma 16. In addition, applying Lemma 17 to Lemma 16 as well, we have

$$V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k) \leq -\left(\frac{\lambda+1}{2}\right) q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (\lambda+1)q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{q\alpha}{2}\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 - \frac{q\beta}{2}\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{(2\lambda+1)\alpha+\beta}{2} qL^2 G_k. \quad (18)$$

If we take the conditional expectation $\mathbb{E}_k$ and apply Lemma 18 (which requires the third inequality of (17) to hold) to (18)

$$\begin{aligned}
&\mathbb{E}_k[V_\lambda(\boldsymbol{z}_0^{k+1})] - V_\lambda(\boldsymbol{z}_0^k) \\
&\leq -\left(\frac{\lambda+1}{2}\right) q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (\lambda+1)q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\
&\quad + \frac{1}{2}\left[q\alpha + 2L^2 q\left(q^2 + \frac{q(q-1)}{n-1}A\right)((2\lambda+1)\alpha+\beta)\,\alpha^2\right]\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\
&\quad \underbrace{- \frac{1}{2}\left[q\beta - 2L^2 q\left(q^2 + \frac{q(q-1)}{n-1}A\right)((2\lambda+1)\alpha+\beta)\,\beta^2\right]}_{C_0}\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \\
&\quad \underbrace{+ \left(\frac{L^2 q^2(q-1)}{n-1}\right)((2\lambda+1)\alpha+\beta)\,(\alpha^2+\beta^2)B}_{C_3}. \qquad (19)
\end{aligned}$$

It is now left to bound terms in (19) using the tools developed so far. First, recall that $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}'\in\mathcal{Y}} f(\boldsymbol{x};\boldsymbol{y}')$. Since $-f(\boldsymbol{x};\boldsymbol{y})$ is μ_2-PŁ in $\boldsymbol{y}$, we have

$$-\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \leq -2\mu_2(\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)). \quad (20)$$

Given any $\boldsymbol{x}$, $\nabla\Phi(\boldsymbol{x}) = \nabla_1 f(\boldsymbol{x};\boldsymbol{y}^*(\boldsymbol{x}))$ for any $\boldsymbol{y}^*(\boldsymbol{x}) \in \arg\max_{\boldsymbol{y}'\in\mathcal{Y}} f(\boldsymbol{x};\boldsymbol{y}')$ by Proposition 9. Besides, $-f(\boldsymbol{x};\cdot)$ satisfies QG condition with constant μ_2 by Proposition 7. Thus, by choosing $\boldsymbol{y}^*(\boldsymbol{x}_0^k)$ to be the projection of $\boldsymbol{y}_0^k$ onto $\arg\max_{\boldsymbol{y}'\in\mathcal{Y}} f(\boldsymbol{x}_0^k;\boldsymbol{y}')$,

$$\left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \leq L^2\left\|\boldsymbol{y}^*(\boldsymbol{x}_0^k) - \boldsymbol{y}_0^k\right\|^2 \leq 2L\kappa_2\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right]. \quad (21)$$

Here, the first inequality applies L-Lipschitz continuity of $\nabla_1 f(\boldsymbol{x}_0^k;\cdot)$, implied by Assumption 1. On top of that, applying the Young's inequality to the term $\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2$,

$$\begin{aligned}
\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 &\leq 2\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + 2\left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\
&\overset{(21)}{\leq} 2\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + 4L\kappa_2\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right] \qquad (22)
\end{aligned}$$

By applying inequalities (20), (21), and (22) to the bound (19), we conclude the proof. □

In Lemma 19, we saw that if step sizes are chosen to satisfy certain conditions, then we can simplify the per-epoch progress a great deal. It is now left to choose appropriate step sizes and parameters (*e.g.*, λ) so as to make sure not only that α and β meet the *small* step size conditions (17) but also that the constants C_0, C_1, C_2, and C_3 are positive.

Lemma 20. *Suppose that Assumptions 1, 2, 3 and 4 hold. Let $\lambda = 4$ and assume that*

$$0 < \beta \leq \frac{1}{6L\sqrt{q^2 + \frac{q(q-1)}{n-1}A}}, \quad \alpha = \frac{\beta}{r}, \quad \textit{where } r \geq 14\kappa_2^2.$$

Then these satisfy all the inequalities (17) and the terms defined in Lemma 19 satisfy

$$C_0 > 0, \quad C_1 > q\alpha, \quad C_2 > L\kappa_2 q\alpha/2, \quad C_3 \geq 0.$$

Consequently, due to the recurrence inequality in Lemma 19, mini-batch simSGDA-RR satisfies, for some numerical constant $c > 0$,

$$\begin{aligned}
\mathbb{E}_k[V_\lambda(\boldsymbol{z}_0^{k+1})] - V_\lambda(\boldsymbol{z}_0^k) &\leq -q\alpha\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 - (L\kappa_2 q\alpha/2)\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right] \\
&\quad + (cr)^3 L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3. \qquad (\star)
\end{aligned}$$

Please note that we mark the recurrence inequality above with a special symbol ($\star$) because this inequality is the exact point where the proofs of Theorems 4 and 5 start to deviate.

Proof. Regardless of $A \geq 0$, we have

$$\beta \leq \frac{1}{6Lq} \quad \text{and} \quad \alpha \leq \frac{1}{6Lqr} \leq \frac{1}{84L\kappa_2^2 q}. \tag{23}$$

This is enough to guarantee that the inequalities (17) hold with $\lambda = 4$. Since $C_0 > C_2/\mu_2$, it is enough to show $C_2 > 0$ to prove that $C_0 > 0$. Applying $\lambda = 4$, $\kappa_2 \geq 1$, and $\beta/\alpha = r \geq 14\kappa_2^2$,

$$\begin{aligned}
\frac{C_1}{q\alpha} &= \frac{3}{2} - 2L^2\left(q^2 + \frac{q(q-1)}{n-1}A\right)(9+r)\,\alpha^2 \\
&\geq \frac{3}{2} - \frac{2}{6^2}\cdot\frac{9+r}{r^2} \geq \frac{3}{2} - \frac{2\cdot 23}{6^2\cdot 14^2} > 1, \\
\frac{C_2}{\mu_2 q\beta} &= 1 - \frac{12\kappa_2^2}{r} - 2L^2\left(q^2 + \frac{q(q-1)}{n-1}A\right)\left(\frac{9}{r}+1\right)\left(\frac{2\kappa_2^2}{r^2}+1\right)\beta^2 \\
&\geq 1 - \frac{12}{14} - \frac{2}{6^2}\left(\frac{9}{14\kappa_2^2}+1\right)\left(\frac{2}{14^2\kappa_2^2}+1\right) \geq \frac{2}{14} - \frac{2\cdot 23\cdot 198}{6^2\cdot 14^3} > \frac{1}{2\cdot 14}.
\end{aligned}$$

Thus, $C_1 > q\alpha$ and

$$C_2 > \frac{\mu_2 q\beta}{2\cdot 14} = \frac{\mu_2 qr\alpha}{2\cdot 14} \geq L\kappa_2 q\alpha/2.$$

Then we conclude the proof by bounding the term C_3. We can already check from the definition that $C_3 \geq 0$. We can upper-bound C_3 by

$$C_3 = \left(\frac{L^2q^2(q-1)}{n-1}\right)(9+r)\,(1+r^2)B\alpha^3 \leq (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3,$$

for some numerical constant $c > 0$. □

C.4 Convergence rates for smooth nonconvex-PŁ problem

In this subsection, we show the convergence bound of general smooth nonconvex-PŁ problems in terms of $\min_{k\in[K]} \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right]$. From the inequality ($\star$) in Lemma 20, we can simply ignore the second term

$$-(L\kappa_2 q\alpha/2)\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right] \leq 0$$

of the right-hand side because $\Phi(\boldsymbol{x}) \geq f(\boldsymbol{x};\boldsymbol{y})$ for any $(\boldsymbol{x};\boldsymbol{y})$. In other words, we may deal with the inequality

$$\mathbb{E}_k[V_\lambda(\boldsymbol{z}_0^{k+1})] - V_\lambda(\boldsymbol{z}_0^k) \leq -q\alpha\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3. \tag{nc-PŁ}$$

Plugging $q = n/b$, we eventually show the convergence rate (Theorem 4). (Recall that b is the size of mini-batches.)

Theorem 21 (Equivalent to Theorem 4, for simSGDA-RR)**.** *Suppose that f satisfies Assumptions 1, 2, 3, and 4 are satisfied. Let $\lambda = 4$. Choose the step sizes α and β by $\alpha = \beta/r$ for some $r \geq 14\kappa_2^2$ and*

$$\beta = \min\left\{\frac{1}{6L\sqrt{q^2 + \frac{q(q-1)}{n-1}A}},\ \frac{1}{c}\left(\frac{V_\lambda(\boldsymbol{z}_0^1)}{L^2q^2(\frac{q-1}{n-1})BK}\right)^{\frac{1}{3}}\right\},$$

for some numerical constant $c > 0$. Then, mini-batch simSGDA-RR satisfies

$$\begin{aligned}
\min_{k\in[K]} \mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] \\
&< \frac{6rLV_\lambda(\boldsymbol{z}_0^1)}{K}\sqrt{1 + \left(\frac{q-1}{n-1}\right)\frac{A}{q}} + 2cr\left(\frac{L^2B\,V_\lambda(\boldsymbol{z}_0^1)^2}{qK^2}\cdot\frac{q-1}{n-1}\right)^{1/3}.
\end{aligned}$$

Proof. To replace the conditional expectations with unconditional expectations, we take expectation to both sides of the inequality (nc-PŁ):

$$\mathbb{E}[V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k)] \leq -q\alpha\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] + (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3.$$

Rearranging the terms and taking a sum from $k=1$ to $k=K$, we have

$$q\alpha\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] \leq \mathbb{E}[V_\lambda(\boldsymbol{z}_0^1) - V_\lambda(\boldsymbol{z}_0^{K+1})] + (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3K.$$

Dividing both sides by $qK\alpha$, we get the following. Note that V_λ is non-negative.

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] \leq \frac{V_\lambda(\boldsymbol{z}_0^1)}{qK\alpha} + (cr)^3L^2\left(\frac{q(q-1)}{n-1}\right)B\alpha^2$$

Since our choice of step sizes implies that

$$\alpha = \min\left\{\frac{1}{6rL\sqrt{q^2+\frac{q(q-1)}{n-1}A}},\ \frac{1}{cr}\left(\frac{V_\lambda(\boldsymbol{z}_0^1)}{L^2Bq^2(\frac{q-1}{n-1})K}\right)^{\frac{1}{3}}\right\}.$$

Using the inequality $\max\{a,b\} \leq a+b$ (for $a,b \geq 0$), we eventually have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2\right] < \frac{6rLV_\lambda(\boldsymbol{z}_0^1)}{K}\sqrt{1+\left(\frac{q-1}{n-1}\right)\frac{A}{q}} + 2cr\left(\frac{L^2B\,V_\lambda(\boldsymbol{z}_0^1)^2}{qK^2}\cdot\frac{q-1}{n-1}\right)^{1/3}.$$

□

C.5 Convergence rates for smooth PŁ(Φ)-PŁ problem

In this subsection, we prove the convergence bound of PŁ(Φ)-PŁ problems in terms of $\mathbb{E}\left[V_\lambda(\boldsymbol{z}_0^{K+1})\right]$.

Unlike the previous subsection, we additionally utilize Assumption 5 stating that $f(\boldsymbol{x};\boldsymbol{y})$ satisfies primal PŁ condition, namely, the primal function $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}'} f(\boldsymbol{x};\boldsymbol{y}')$ is a μ_1-PŁ function. With this assumption, we yield another recurrence inequality from the inequality ($\star$). We note that it uses the μ_1-PŁ condition for Φ ($\because$ Proposition 10) but not necessarily for $f(\cdot;\boldsymbol{y})$.

Lemma 22. *Suppose that f satisfies Assumptions 1, 2, 3, 4, and 5. Then, with the same choice of $\lambda = 4$ and the same condition of the step sizes α and β as in Lemma 20, the mini-batch simSGDA-RR satisfies that, for some numerical constant $c > 0$,*

$$\mathbb{E}_k[V_\lambda(\boldsymbol{z}_0^{k+1})] \leq (1-\mu_1q\alpha/2)V_\lambda(\boldsymbol{z}_0^k) + (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3. \tag{2PŁ}$$

Proof. Since the primal function Φ is a μ_1-PŁ function,

$$-\left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 \leq -2\mu_1\left[\Phi(\boldsymbol{x}_0^k) - \Phi^*\right].$$

Also, since $\mu_1 \leq L$ and $\kappa_2 \geq 1$, we know that $-L\kappa_2 \leq -\mu_1$. Applying these to the inequality ($\star$), we have

$$\begin{aligned}
&\mathbb{E}_k\left[V_\lambda(\boldsymbol{z}_0^{k+1})\right] - V_\lambda(\boldsymbol{z}_0^k)\\
&\leq -(2\mu_1q\alpha/\lambda)\cdot\lambda\left[\Phi(\boldsymbol{x}_0^k) - \Phi^*\right] - (\mu_1q\alpha/2)\left[\Phi(\boldsymbol{x}_0^k) - f(\boldsymbol{z}_0^k)\right] + (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3\\
&= -(\mu_1q\alpha/2)\cdot V_\lambda(\boldsymbol{z}_0^k) + (cr)^3L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3,
\end{aligned}$$

since $\lambda = 4$. By re-arranging the terms, we conclude the proof. □

Of course, the multiplier $1 - \mu_1 q\alpha/2$ have a value between 0 and 1. To see why, note that from Equation (23),

$$0 < \mu_1 q\alpha/2 \leq \mu_1 q \cdot \frac{1}{2 \cdot 84 L\kappa_2^2 q} = \frac{1}{168\kappa_1\kappa_2^2} < 1.$$

Theorem 23 (Equivalent to Theorem 5, for simSGDA-RR)**.** *Assume that f satisfies Assumptions 1, 2, 3, 4, and 5. Let $\lambda = 4$. Choose the step sizes by $\alpha = \beta/r$ for some $r \geq 14\kappa_2^2$ and*

$$\beta = \min\left\{\frac{1}{6L\sqrt{q^2 + \frac{q(q-1)}{n-1}A}}, \frac{2r}{\mu_1 qK}\max\left\{1, \log\left(\frac{V_\lambda(\boldsymbol{z}_0^1)\mu_1 qK^2}{8(cr)^3\kappa_1^2\left(\frac{q-1}{n-1}\right)B}\right)\right\}\right\},$$

for some numerical constant $c > 0$. Then, mini-batch simSGDA-RR satisfies

$$\mathbb{E}[V_\lambda(\boldsymbol{z}_n^K)] \leq \mathcal{O}\left(V_\lambda(\boldsymbol{z}_0^1)\cdot\exp\left(-\frac{K}{12\kappa_1 r\sqrt{1 + \left(\frac{q-1}{n-1}\right)\frac{A}{q}}}\right)\right) + \tilde{\mathcal{O}}\left(\frac{\kappa_1^2 r^3 B}{\mu_1 qK^2}\right)\cdot\frac{q-1}{n-1}.$$

Proof. To replace the conditional expectations with unconditional expectations, we take expectation to both sides of the inequality (2PŁ):

$$\mathbb{E}\left[V_\lambda(\boldsymbol{z}_0^{k+1})\right] \leq (1 - \mu_1 q\alpha/2)\mathbb{E}\left[V_\lambda(\boldsymbol{z}_0^k)\right] + (cr)^3 L^2\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3.$$

Unrolling the recurrence inequality (Proposition 15) and using the facts $\beta = 14\kappa_2^2\alpha$, we have

$$\begin{aligned}\mathbb{E}[V_\lambda(\boldsymbol{z}_n^K)] &\leq (1 - \mu_1 q\alpha/2)^K V_\lambda(\boldsymbol{z}_0^1) + \frac{2\cdot(cr)^3 L^2}{\mu_1 q\alpha}\left(\frac{q^2(q-1)}{n-1}\right)B\alpha^3 \\ &\leq \exp(-\mu_1 qK\alpha/2)V_\lambda(\boldsymbol{z}_0^1) + 2(cr)^3\mu_1\kappa_1^2\left(\frac{q(q-1)}{n-1}\right)B\alpha^2. \qquad (24)\end{aligned}$$

Note that, in the inequality above, the second term of the right hand side becomes zero when $q = 1$. In that case, we can prove exponential decay of $\mathbb{E}[V_\lambda(\boldsymbol{z}_0^k)]$. Thus, we simply assume $q > 1$ hereafter.

Case 1: If K is as large as

$$K > \frac{\kappa_1 r^{3/2}}{\sqrt{\mu_1}}\cdot\sqrt{\frac{8c^3 eB}{V_\lambda(\boldsymbol{z}_0^1)\,q}\left(\frac{q-1}{n-1}\right)}, \quad (e = \exp(1))$$

we have a step size α as

$$\alpha = \min\left\{\frac{1}{6Lr\sqrt{q^2 + \frac{q(q-1)}{n-1}A}}, \frac{2}{\mu_1 qK}\log(\clubsuit)\right\}, \quad \text{where } \clubsuit = \frac{V_\lambda(\boldsymbol{z}_0^1)\mu_1 qK^2}{8(cr)^3\kappa_1^2\kappa_2^6\left(\frac{q-1}{n-1}\right)B}.$$

Due to the lower bound of epoch size K, the fraction $\clubsuit$ inside the log factor is indeed greater than $e > 1$, which guarantees the step size is positive. Putting this to the inequality (24) and using the fact that $\max\{a, b\} \leq a + b$ (for $a, b \geq 0$), we eventually have

$$\begin{aligned}&\mathbb{E}\left[V_\lambda(\boldsymbol{z}_n^K)\right] \\ &\leq V_\lambda(\boldsymbol{z}_0^1)\cdot\exp\left(-\frac{K}{12\kappa_1 r\sqrt{1 + \left(\frac{q-1}{n-1}\right)\frac{A}{q}}}\right) + \frac{2\cdot 8(cr)^3\kappa_1^2 B}{\mu_1 qK^2}\left(\frac{q-1}{n-1}\right)\left[1 + \log^2(\clubsuit)\right] \\ &= V_\lambda(\boldsymbol{z}_0^1)\cdot\exp\left(-\frac{K}{12\kappa_1 r\sqrt{1 + \left(\frac{q-1}{n-1}\right)\frac{A}{q}}}\right) + \tilde{\mathcal{O}}\left(\frac{\kappa_1^2 r^3 B}{\mu_1 qK^2}\right)\cdot\frac{q-1}{n-1}.\end{aligned}$$

Case 2: Otherwise, the log factor might have a negative value when K is too small. However, in this case, we have

$$V_\lambda(\boldsymbol{z}_0^1) \leq \frac{8(cr)^3 e\kappa_1^2 B}{\mu_1 q K^2} \cdot \frac{q-1}{n-1}\,; \quad \alpha = \min\left\{ \frac{1}{84L\kappa_2^2\sqrt{q^2 + \frac{q(q-1)}{n-1}A}},\ \frac{2}{\mu_1 q K} \right\}.$$

Putting these to the inequality (24), we have

$$\begin{aligned}
\mathbb{E}\left[V_\lambda(\boldsymbol{z}_n^K)\right] &\leq \frac{8(cr)^3 e\kappa_1^2 B}{\mu_1 q K^2}\left(\frac{q-1}{n-1}\right)\left[\exp(-\mu_1 q K\alpha/2) + \frac{1}{e}\cdot(\mu_1 q K\alpha/2)^2\right] \\
&\leq \frac{8(cr)^3 e\kappa_1^2 B}{\mu_1 q K^2}\left(\frac{q-1}{n-1}\right) = \mathcal{O}\left(\frac{\kappa_1^2 r^3 B}{\mu_1 q K^2}\right)\cdot\frac{q-1}{n-1}.
\end{aligned}$$

The inequlity in the last line is due to the fact that $e^{-t} + t^2/e \leq 1$ for each $t \in (0,1]$, and that $\mu_1 q K\alpha/2 \in (0,1]$.

In either case, the conclusion does not change:

$$\mathbb{E}\left[V_\lambda(\boldsymbol{z}_n^K)\right] \leq \mathcal{O}\left(V_\lambda(\boldsymbol{z}_0^1)\cdot\exp\left(-\frac{K}{12\kappa_1 r\sqrt{1+\left(\frac{q-1}{n-1}\right)\frac{A}{q}}}\right)\right) + \tilde{\mathcal{O}}\left(\frac{\kappa_1^2 r^3 B}{\mu_1 q K^2}\right)\cdot\frac{q-1}{n-1}.$$

□

D PROOFS FOR (MINI-BATCH) ALTERNATING SGDA-RR: FOCUSING ON CHANGES IN THE PROOF

In this appendix, we prove the same convergence rates for altSGDA-RR as the simultaneous update counterpart. Since most of the steps in the proof is similar to those in Appendix C, we only describe which steps change in the proof.

D.1 EPOCH-WISE REPRESENTATIONS AND BOUNDING NOISE TERMS

To analyze altSGDA-RR, we modify the notation for epoch-wise updates. The only change is that an update $\boldsymbol{y}_t^k \mapsto \boldsymbol{y}_{t+1}^k$ uses $\boldsymbol{x}_{t+1}^k$ instead of $\boldsymbol{x}_t^k$. Hence, the definition of $\boldsymbol{h}^k$ should be modified. Recall that

$$\boldsymbol{g}_t^k(\boldsymbol{z}) := \frac{1}{b}\sum_{i\in\mathcal{B}_t^k} \nabla_1 f_i(\boldsymbol{z}), \quad \boldsymbol{h}_t^k(\boldsymbol{z}) := \frac{1}{b}\sum_{i\in\mathcal{B}_t^k} \nabla_2 f_i(\boldsymbol{z}),$$

where $\mathcal{B}_t^k$ is a mini-batch of size b formed at iteration t of epoch k. Then, at epoch k, by re-definition of $\boldsymbol{h}^k$,

$$\boldsymbol{g}^k := \frac{1}{q}\sum_{t=1}^{q} \boldsymbol{g}_t^k(\boldsymbol{x}_{t-1}^k;\boldsymbol{y}_{t-1}^k), \quad \boldsymbol{h}^k := \frac{1}{q}\sum_{t=1}^{q} \boldsymbol{h}_t^k(\boldsymbol{x}_t^k;\boldsymbol{y}_{t-1}^k).$$
$$\boldsymbol{x}_0^{k+1} = \boldsymbol{x}_0^k - q\alpha\boldsymbol{g}^k, \quad \boldsymbol{y}_0^{k+1} = \boldsymbol{y}_0^k + q\beta\boldsymbol{h}^k. \tag{altSGDA-RR}$$

We still approximate this epoch-wise update rule to a full-batch simultaneous GDA update (≈simGDA) with step sizes $q\alpha$ and $q\beta$. Again, we control the "noise" terms $\left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2$ and $\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2$ not to be large. Because of the modification of $\boldsymbol{h}^k$, we have a different result for $\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2$ as follows.

Lemma 24. *For mini-batch altSGDA-RR, recall that*

$$G_k := \frac{1}{q}\sum_{t=1}^{q} \left\|\boldsymbol{z}_{t-1}^k - \boldsymbol{z}_0^k\right\|^2.$$

If we have Assumption 1, then we have

$$\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \le L^2 G_k + L^2 q\alpha^2 \left\|\boldsymbol{g}^k\right\|^2, \quad \textit{whereas} \quad \left\|\boldsymbol{g}^k - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \le L^2 G_k. \tag{25}$$

Proof. Because of L-Lipschitz continuity of $\boldsymbol{h}_t^k(\cdot;\cdot)$,

$$\begin{aligned}
\left\|\boldsymbol{h}^k - \nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 &= \left\|\frac{1}{q}\sum_{t=1}^{q}\left[\boldsymbol{h}_t^k(\boldsymbol{x}_t^k;\boldsymbol{y}_{t-1}^k) - \boldsymbol{h}_t^k(\boldsymbol{x}_0^k;\boldsymbol{y}_0^k)\right]\right\|^2 \\
&\le \frac{1}{q}\sum_{t=1}^{q}\left\|\boldsymbol{h}_t^k(\boldsymbol{x}_t^k;\boldsymbol{y}_{t-1}^k) - \boldsymbol{h}_t^k(\boldsymbol{x}_0^k;\boldsymbol{y}_0^k)\right\|^2 \\
&\le \frac{L^2}{q}\sum_{t=1}^{q}\left\|\boldsymbol{z}_{t-1}^k - \boldsymbol{z}_0^k\right\|^2 + \frac{L^2}{q}\left\|\boldsymbol{x}_q^k - \boldsymbol{x}_0^k\right\|^2 = L^2 G_k + L^2 q\alpha^2\left\|\boldsymbol{g}^k\right\|^2.
\end{aligned}$$

The last ineqaulity holds because $\boldsymbol{x}_q^k = \boldsymbol{x}_0^{k+1}$. □

D.2 BOUNDING NOISE TERMS: A BIT DIFFERENT PROOF OF LEMMA 18

We notice that the same result as Lemma 18 holds not only for simultaneous updates but also alternating updates, even though it is not very straightforward. We need to reflect the changes from the previous subsection. That is, we have to be careful when we expand the term $\left\|\boldsymbol{y}_t^k - \boldsymbol{y}_0^k\right\|^2$

($0 \le t \le q-1$). Unlike the inequality (12) (in the original proof), we have

$$
\begin{aligned}
&\left\|\boldsymbol{y}_t^k - \boldsymbol{y}_0^k\right\|^2 = \beta^2 t^2 \left\| \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{x}_j^k; \boldsymbol{y}_{j-1}^k) \right\|^2 \\
&\le 3\beta^2 t^2 \left[\left\| \frac{1}{t}\sum_{j=1}^{t} \left[\boldsymbol{h}_j^k(\boldsymbol{x}_j^k; \boldsymbol{y}_{j-1}^k) - \boldsymbol{h}_j^k(\boldsymbol{z}_0^k)\right] \right\|^2 + \left\| \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \right] \\
&\le 3\beta^2 t \sum_{j=1}^{t} \left\|\boldsymbol{h}_j^k(\boldsymbol{x}_j^k; \boldsymbol{y}_{j-1}^k) - \boldsymbol{h}_j^k(\boldsymbol{z}_0^k)\right\|^2 + 3\beta^2 t^2 \left[\left\| \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \right] \\
&\le 3\beta^2 L^2 t \left(\sum_{j=1}^{t} \left\|\boldsymbol{z}_{j-1}^k - \boldsymbol{z}_0^k\right\|^2 + \left\|\boldsymbol{x}_t^k - \boldsymbol{x}_0^k\right\|^2 \right) + 3\beta^2 t^2 \left[\left\| \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \right] \\
&\le 3\beta^2 L^2 t \sum_{j=1}^{t} \left\|\boldsymbol{z}_j^k - \boldsymbol{z}_0^k\right\|^2 + 3\beta^2 t^2 \left[\left\| \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \right] \\
&\le 3\beta^2 L^2 t \cdot qG_k + 3\beta^2 t^2 \left[\left\| \frac{1}{t}\sum_{j=1}^{t} \boldsymbol{h}_j^k(\boldsymbol{z}_0^k) - \nabla_2 f(\boldsymbol{z}_0^k) \right\|^2 + \left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 \right].
\end{aligned}
$$

The second and third inequality holds by Jensen's inequality, and the last inequality holds because $t \le q-1$. The resulting upper bound is identical to the inequality (13). Proving this inequality above suffices to show that the conclusion of Lemma 18 also holds for altSGDA-RR, because we eventually take an average along $0 \le t \le q-1$ and the other steps in the proof does not utilize the "order" (either simultaneous or alternating) of updates.

D.3 Recurrence inequalities for general smooth nonconvex-PŁ objective

In the proof for simSGDA-RR, we applied Lemma 16, Lemma 18, and the "small-step-size" assumptions (three inequalities in (17)) to deduce Lemma 19. However, due to Lemma 24 that we obtained for altSGDA-RR, we need slightly different assumptions on step sizes rather than (17).

Fortunately, we notice that the Lemma 16 also holds for altSGDA-RR, with a modified version of $\boldsymbol{h}^k$. This is because the proof of the lemma does not utilize step-wise updates, while the discrepancy between simultaneous and alternating updates only appear in the step-wise updates. Thus, we have the same result as Lemma 19.

Lemma 25. *Suppose that Assumptions 1, 2, 3, and 4 hold. Modify the inequalities (17) (from Lemma 19) by*

$$
\lambda - \{(\lambda+1)(\kappa_2+1)+1\} Lq\alpha - L^2 q\alpha\beta \ge 0, \quad \beta \le \frac{1}{qL}, \quad \alpha^2+\beta^2 \le \frac{1}{3q(q-1)L^2}. \tag{26}
$$

(In fact, only the first one is different.) Then, the result of Lemma 19 still holds for mini-batch altSGDA-RR.

Proof. We first apply Lemma 24 to the general bound resulted from Lemma 16:

$$
\begin{aligned}
&V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k) \\
&\le -\left(\frac{\lambda+1}{2}\right) q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (\lambda+1)q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 \\
&\quad + \frac{q\alpha}{2}\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 - \frac{q\beta}{2}\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{(2\lambda+1)\alpha+\beta}{2} qL^2 G_k \\
&\quad - \left[\lambda - \{(\lambda+1)(\kappa_2+1)+1\} Lq\alpha - L^2 q\alpha\beta\right] \frac{q\alpha}{2}\left\|\boldsymbol{g}^k\right\|^2 - (1-Lq\beta)\frac{q\beta}{2}\left\|\boldsymbol{h}^k\right\|^2.
\end{aligned} \tag{27}
$$

Hence, the first two inequalities of (26) eliminates the last two terms on the right side of the inequality (27) above:

$$V_\lambda(\boldsymbol{z}_0^{k+1}) - V_\lambda(\boldsymbol{z}_0^k) \leq -\left(\frac{\lambda+1}{2}\right) q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k)\right\|^2 + (\lambda+1)q\alpha \left\|\nabla\Phi(\boldsymbol{x}_0^k) - \nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{q\alpha}{2}\left\|\nabla_1 f(\boldsymbol{z}_0^k)\right\|^2 - \frac{q\beta}{2}\left\|\nabla_2 f(\boldsymbol{z}_0^k)\right\|^2 + \frac{(2\lambda+1)\alpha+\beta}{2} qL^2 G_k.$$

This is identical to the inequality (18) in the proof of Lemma 19. From this point on, the rest of proof is exactly identical to Lemma 19. □

Lemma 25 establishes that altSGDA-RR also satisfies a concise bound on the expected per-epoch change of V_λ, albeit under a slightly different set of assumptions (26) on step sizes. Using this result, we can prove the convergence rates for altSGDA-RR that are exactly the same as simSGDA-RR.

D.4 Small step size assumptions

It is left to show an altSGDA-RR counterpart for Lemma 20 which establishes the general recurrence inequality ($\star$). In fact, the same choice of step sizes as simSGDA-RR, namely

$$0 < \beta \leq \frac{1}{6L\sqrt{q^2 + \frac{q(q-1)}{n-1}A}} \quad \text{and} \quad \alpha = \frac{\beta}{r} \quad \text{where } r \geq 14\kappa_2^2,$$

actually meets the newly introduced conditions (26). Among three inequalities, the only one that needs to be checked is

$$\lambda - \{(\lambda+1)(\kappa_2+1)+1\} Lq\alpha - L^2 q\alpha\beta > 0.$$

Note that, regardless of $A \geq 0$,

$$\beta \leq \frac{1}{6Lq} \quad \text{and} \quad \alpha \leq \frac{1}{6Lqr} \leq \frac{1}{84L\kappa_2^2 q}$$

In this case,

$$\begin{aligned}
&\lambda - \{(\lambda+1)(\kappa_2+1)+1\} Lq\alpha - L^2 q\alpha\beta \\
&\geq 4 - (11\kappa_2 + L\beta)Lq\alpha \geq 4 - \left(11\kappa_2 + \frac{1}{6}\right)\cdot\frac{1}{84\kappa_2^2} > 0.
\end{aligned}$$

Therefore, there is no need to modify our choices of λ and the step sizes α, β for analysis of altSGDA-RR, and the rest of the proof for simSGDA-RR goes through.

E PROOFS FOR LOWER BOUND OF TWO-TIME-SCALE GDA

In this appendix, we illustrate a comprehensive lower bound for full-batch two-time-scale GDA which is specific to the choice of step size ratio (Theorem 3). The proof of Theorem 3 is at the end of this appendix.

Before we start the proof, we define a class of smooth strongly-convex-strongly concave functions.

Definition 4. *Let $\mathcal{F}(L, \mu_1, \mu_2)$ be the class of functions $f(\boldsymbol{x}; \boldsymbol{y})$ with two arguments $\boldsymbol{x}$ and $\boldsymbol{y}$ of any dimension, which is L-smooth, μ_1-strongly-convex in $\boldsymbol{x}$, and μ_2-strongly-concave in $\boldsymbol{y}$. Let $\kappa_1 = L/\mu_1 \geq 1$ and $\kappa_2 = L/\mu_2 \geq 1$ be condition numbers of the function class. Denote the (unique) saddle (or, global minimax) point by $\boldsymbol{z}^* = (\boldsymbol{x}^*; \boldsymbol{y}^*)$.*

We restate and prove the Theorem 3 for reader's convenience.

Theorem 26 (Restatement of Theorem 3)**.** *Suppose $\kappa_1 \geq c$ and $\kappa_2 \geq c$ for some constant $c > 1$. Then, for each step size ratio $r > 0$, there exists a function $f \in \mathcal{F}(L, \mu_1, \mu_2)$ for which two-time-scale GDA with any step sizes α and β of ratio $r = \beta/\alpha$ requires*

$$K = \begin{cases} \Omega\left(\kappa_1 r \log(1/\varepsilon)\right), & \text{if } r \geq \kappa_2/c, \\ \Omega\left(\kappa_1 \kappa_2 \log(1/\varepsilon)\right), & \text{if } c/\kappa_1 \leq r \leq \kappa_2/c, \\ \Omega((\kappa_2/r)\log(1/\varepsilon)), & \text{if } 0 < r \leq c/\kappa_1 \end{cases}$$

iterations to achieve either $\|\boldsymbol{z}_k - \boldsymbol{z}^*\|^2 \leq \varepsilon^2$ or $V_\lambda(\boldsymbol{z}_K) \leq \varepsilon^2$.

Proof. The proof is done in case by case, constructing a worst-case function for each of 4 different regimes of step size ratio r: *(1)* $\mu_1/\mu_2 \leq r \leq \kappa_2/c$, *(2)* $c/\kappa_1 \leq r \leq \mu_1/\mu_2$, *(3)* $r \geq \kappa_2/c$, and *(4)* $0 < r \leq c/\kappa_1$. Readers might notice the similarities of the proofs for *(1)*↔*(2)* and *(3)*↔*(4)*.

<u>Case 1.</u> ($\mu_1/\mu_2 \leq r \leq \kappa_2/c$). Consider

$$f^{(1)}(v, x; y) := \frac{\mu_1}{2}v^2 + \frac{r\mu_2}{2}x^2 - \frac{\mu_2}{2}y^2 + \ell xy,$$

where $\ell^2 = L^2 - r\mu_2^2 - L\mu_2|r-1| \geq 0$. Applying Proposition 28, it can be shown that $f^{(1)} \in \mathcal{F}(L, \mu_1, \mu_2)$. Also, $\boldsymbol{z}^* = (0, 0; 0)$ is its unique saddle point. Note that, the GDA on $f^{(1)}$ can be written as

$$v_{t+1} = \left(1 - \frac{\beta\mu_1}{r}\right)v_t, \quad \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1-\beta\mu_2 & -\beta\ell/r \\ \beta\ell & 1-\beta\mu_2 \end{bmatrix}}_{\boldsymbol{A}} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \boldsymbol{A}\begin{bmatrix} x_t \\ y_t \end{bmatrix}.$$

Also, the eigenvalues τ of $\boldsymbol{A}$ is

$$\begin{aligned} \tau &= 1 - \beta\mu_2 \pm \sqrt{(1-\beta\mu_2)^2 - ((1-\beta\mu_2)^2 + \beta^2\ell^2/r)} \\ &= 1 - \beta\mu_2 \pm \frac{\beta\ell}{\sqrt{r}}\sqrt{-1}. \end{aligned}$$

The spectral radius (*i.e.*, maximum absolute eigenvalue) is

$$\rho(\boldsymbol{A}) = \sqrt{(1-\beta\mu_2)^2 + \beta^2\ell^2/r}.$$

Since the eigenvalues are complex conjugates of each other (the magnitudes are the same), both eigenvalues have magnitude $\rho(\boldsymbol{A})$. Then, by Proposition 27, $\rho(\boldsymbol{A}) < 1$ is necessary for convergence. To this end, we need $\beta > 0$ satisfying $\beta < 2\mu_2 r/(r\mu_2^2 + \ell^2)$.

To guarantee $\|(v_k, x_k; y_k) - (0, 0; 0)\|^2 \leq \varepsilon^2$, we need a large enough k to have $v_k^2 \leq \mathcal{O}(\varepsilon^2)$. Such a k is required to be at least $\Omega\left(\frac{r}{\beta\mu_1}\log(1/\varepsilon)\right)$. Now note that, since $\mu_1/\mu_2 \leq r \leq \kappa_2/c$ and $\kappa_2 \geq c$,

$$\frac{1}{\beta} > \frac{r\mu_2^2 + \ell^2}{2\mu_2 r} = \frac{L^2 - L\mu_2|r-1|}{2\mu_2 r} = \frac{L^2}{2\mu_2 r}\left(1 - \frac{|r-1|}{\kappa_2}\right) \geq \frac{L^2}{2\mu_2 r}\left(1 - \frac{1}{c}\right).$$

The last inequality is true by minimizing $\left(1 - \frac{|r-1|}{\kappa_2}\right)$ for $r \in [\mu_1/\mu_2, \kappa_2/c]$. If $r \geq 1$, it has smaller value when r is larger: by taking $r = \kappa_2/c$, we have $1 - \frac{\kappa_2/c - 1}{\kappa_2} \geq 1 - \frac{1}{c}$. Otherwise ($r < 1$), which

is possible only when $\mu_1 < \mu_2$, the term has smaller value when r is smaller: by taking $r = \mu_1/\mu_2$, we have $1+\frac{\mu_1/\mu_2-1}{\kappa_2} = 1+\frac{\mu_1-\mu_2}{L} \geq 1-\frac{1}{\kappa_2} \geq 1-\frac{1}{c}$. Thus, we eventually need $\Omega\left(\frac{L^2}{\mu_1\mu_2}\log(1/\varepsilon)\right)$ iterations.

Case 2. ($c/\kappa_1 \leq r \leq \mu_1/\mu_2$). Consider

$$f^{(2)}(x;y,w) := \frac{\mu_1}{2}x^2 - \frac{\mu_1}{2r}y^2 + \tilde{\ell}xy - \frac{\mu_2}{2}w^2,$$

where $\tilde{\ell}^2 = L^2 - \mu_1^2/r - L\mu_1|1-1/r| \geq 0$. Applying Proposition 28, it can be shown that $f^{(2)} \in \mathcal{F}(L,\mu_1,\mu_2)$, and $\boldsymbol{z}^* = (0;0,0)$ is its unique saddle point. Note that, the GDA on $f^{(2)}$ can be written as

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1-\beta\mu_1/r & -\beta\ell/r \\ \beta\ell & 1-\beta\mu_1/r \end{bmatrix}}_{\boldsymbol{B}} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \boldsymbol{B}\begin{bmatrix} x_t \\ y_t \end{bmatrix}, \quad w_{t+1} = (1-\beta\mu_2)\, w_t.$$

Also, the eigenvalues τ of $\boldsymbol{B}$ is

$$\begin{aligned} \tau &= 1-\beta\mu_1/r \pm \sqrt{(1-\beta\mu_1/r)^2 - ((1-\beta\mu_1/r)^2 + \beta^2\ell^2/r)} \\ &= 1 - \frac{\beta\mu_1}{r} \pm \frac{\beta\ell}{\sqrt{r}}\sqrt{-1}. \end{aligned}$$

The spectral radius is

$$\rho(\boldsymbol{B}) = \sqrt{(1-\beta\mu_1/r)^2 + \beta^2\ell^2/r}.$$

Since the eigenvalues are complex conjugates of each other (the magnitudes are the same), both eigenvalues have magnitude $\rho(\boldsymbol{A})$. Then, by Proposition 27, $\rho(\boldsymbol{A}) < 1$ is necessary for convergence. To this end, we need $\beta > 0$ satisfying $\beta < 2\mu_1/(\mu_1^2/r + \ell^2)$.

To guarantee $\|(x_k;y_k,w_k) - (0;0,0)\|^2 \leq \varepsilon^2$, we need a large enough k to have $w_k^2 \leq \mathcal{O}(\varepsilon^2)$. Such a k is required to be at least $\Omega\left(\frac{1}{\beta\mu_2}\log(1/\varepsilon)\right)$. Now note that, since $c/\kappa_1 \leq r \leq \mu_1/\mu_2$ and $\kappa_1 \geq c$,

$$\frac{1}{\beta} > \frac{\mu_1^2/r + \ell^2}{2\mu_1} = \frac{L^2 - L\mu_1|1-1/r|}{2\mu_1} = \frac{L^2}{2\mu_1}\left(1 - \frac{|1-1/r|}{\kappa_1}\right) \geq \frac{L^2}{2\mu_1}\left(1-\frac{1}{c}\right).$$

The last inequality is true by minimizing $\left(1-\frac{|1-1/r|}{\kappa_1}\right)$ for $r \in [c/\kappa_1, \mu_1/\mu_2]$. If $1 > 1/r$, which is possible only when $\mu_1 > \mu_2$, it has smaller value when r is larger: by taking $r = \mu_1/\mu_2$, we have $1 - \frac{1-\mu_2/\mu_1}{\kappa_1} = 1 - \frac{\mu_1-\mu_2}{L} \geq 1 - \frac{1}{\kappa_1} \geq 1 - \frac{1}{c}$. Otherwise ($1 < 1/r$), the term has smaller value when r is smaller: by taking $r = c/\kappa_1$, we have $1 + \frac{1-\kappa_1/c}{\kappa_1} \geq 1 - \frac{1}{c}$. Thus, we eventually need $\Omega\left(\frac{L^2}{\mu_1\mu_2}\log(1/\varepsilon)\right)$ iterations.

Case 3. ($r \geq \kappa_2/c$). Consider $f^{(3)}(x;y) = \frac{\mu_1}{2}x^2 - \frac{L}{2}y^2$. Clearly, $f^{(3)} \in \mathcal{F}(L,\mu_1,L) \subset \mathcal{F}(L,\mu_1,\mu_2)$ and $\boldsymbol{z}^* = (0,0)$ is its unique saddle point. The GDA on $f^{(3)}$ can be written as

$$x_{k+1} = \left(1 - \frac{\beta\mu_1}{r}\right)x_k, \quad y_{k+1} = (1-\beta L)\, y_k.$$

To guarantee $\|(x_k;y_k) - (0,0)\|^2 \leq \varepsilon^2$, we need a large enough k to have $x_k^2 \leq \mathcal{O}(\varepsilon^2)$. Such a k is required to be at least $\Omega\left(\frac{r}{\beta\mu_1}\log(1/\varepsilon)\right)$. Also, we need $\beta < 2/L$ to guarantee $y_k \to 0$ (*i.e.*, otherwise, it diverges). Combining these facts, we eventually need $\Omega\left(\frac{Lr}{\mu_1}\log(1/\varepsilon)\right)$ iterations.

Case 4. ($0 < r \leq c/\kappa_1$). Consider $f^{(4)}(x;y) = \frac{L}{2}x^2 - \frac{\mu_2}{2}y^2$. Clearly, $f^{(4)} \in \mathcal{F}(L,L,\mu_2) \subset \mathcal{F}(L,\mu_1,\mu_2)$ and $\boldsymbol{z}^* = (0,0)$ is its unique saddle point. The GDA on $f^{(4)}$ can be written as

$$x_{k+1} = \left(1 - \frac{\beta L}{r}\right)x_k, \quad y_{k+1} = (1-\beta\mu_2)\, y_k.$$

To guarantee $\|(x_k; y_k) - (0,0)\|^2 \le \varepsilon^2$, we need a large enough k to have $y_k^2 \le \mathcal{O}(\varepsilon^2)$. Such a k is required to be at least $\Omega\left(\frac{1}{\beta\mu_2}\log(1/\varepsilon)\right)$. Also, we need $\beta < 2r/L$ to guarantee $x_k \to 0$ (*i.e.*, otherwise, it diverges). Combining these facts, we eventually need $\Omega\left(\frac{L}{r\mu_2}\log(1/\varepsilon)\right)$ iterations.

Lastly, we note that the lower iteration complexity bound in terms of the potential function V_λ is equivalent to the complexity in terms of squared distance norm from the (unique) saddle point $\boldsymbol{z}^*$, up to constant factors. This is proved in Lemma 29 that we defer its proof. □

Here are the postponed proofs from the proof above.

Proposition 27. *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 to $\boldsymbol{v}_k \to \boldsymbol{0}$ if and only if the spectral radius* (i.e., *maximum absolute eigenvalue) of $\rho(\boldsymbol{A})$ of $\boldsymbol{A}$ is less than 1. Furthermore, its convergence speed is characterized by $\mathcal{O}((\rho(\boldsymbol{A})+\varepsilon)^k)$ for any (arbitrarily small) $\varepsilon > 0$.*

Proof. See Horn & Johnson (2012, Theorem 5.6.10-12). □

Proposition 28. *Let μ_1, μ_2, and L be positive numbers such that $L \ge \max\{\mu_1, \mu_2\}$. Consider a quadratic function f on $\mathbb{R}\times\mathbb{R}$ defined by*

$$f(x;y) = \frac{\mu_1}{2}x^2 - \frac{\mu_2}{2}y^2 + \ell xy, \quad \textit{where } \ell^2 \le L^2 - \mu_1\mu_2 - L|\mu_1 - \mu_2|.$$

Then, $f \in \mathcal{F}(L, \mu_1, \mu_2)$, and its unique saddle point is $\boldsymbol{z}^* = (0,0)$.

For example, if $\mu_1 \ge \mu_2$, $\ell^2 = (L-\mu_1)(L+\mu_2)$ is enough to guarantee L-smoothness.

Proof. The strong-convex-strong-concavity is trivially true. Note that the gradient and hessian of f is

$$\nabla f(x;y) = \boldsymbol{H}[x \;\; y]^\top, \quad \boldsymbol{H} = \begin{bmatrix}\mu_1 & \ell \\ \ell & -\mu_2\end{bmatrix}.$$

Since $\boldsymbol{H}$ is a non-singular matrix, f has a unique stationary point at origin. By Proposition 11, it is also a unique saddle & global minimax point.

For any two distinct points $\boldsymbol{z}_1 = (x_1; y_1)$ and $\boldsymbol{z}_2 = (x_2; y_2)$ in $\mathbb{R}\times\mathbb{R}$,

$$\frac{\|\nabla f(\boldsymbol{z}_1) - \nabla f(\boldsymbol{z}_2)\|}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|} = \frac{\|\boldsymbol{H}(\boldsymbol{z}_1 - \boldsymbol{z}_2)\|}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|} \le \|\boldsymbol{H}\|_2,$$

where $\|\boldsymbol{H}\|_2$ is spectral norm (*i.e.*, maximum singular value) of $\boldsymbol{H}$. We would like to show that $\|\boldsymbol{H}\|_2 \le L$. To this end, it is enough to verify the following two inequalities:

$$\det(L^2\boldsymbol{I} - \boldsymbol{H}\boldsymbol{H}^\top) = L^4 - (\mu_1^2 + \mu_2^2 + 2\ell^2)L^2 + (\mu_1\mu_2 + \ell^2)^2 \ge 0,$$
$$\operatorname{trace}(\boldsymbol{H}\boldsymbol{H}^\top)/2 = (\mu_1^2 + \mu_2^2)/2 + \ell^2 \le L^2.$$

This is because the characteristic polynomial of $\boldsymbol{H}\boldsymbol{H}^\top$, or $\det(\omega\boldsymbol{I} - \boldsymbol{H}\boldsymbol{H}^\top)$, is a quadratic polynomial of ω, and its maximum root should not be greater than L^2. Let $\ell^2 = L^2 - \mu_1\mu_2 + a$ for some $a \in \mathbb{R}$. Plugging this ℓ^2 into both inequalities above, we get

$$a^2 - (\mu_1 - \mu_2)^2 L^2 \ge 0 \quad \text{and} \quad a \le -(\mu_1 - \mu_2)^2/2,$$

respectively. One can check that $a = -L|\mu_1 - \mu_2|$ is the largest possible a satisfying both inequalities above. This proves the proposition. □

Subsequently, we show that if our convergence rate is exponential, then the iteration complexity in terms of $\|\boldsymbol{z} - \boldsymbol{z}^*\|^2$ is equivalent to that in terms of $V_\lambda(\boldsymbol{z}) = \lambda[\Phi(\boldsymbol{x}) - \Phi^*] + [\Phi(\boldsymbol{x}) - f(\boldsymbol{z})]$ for PŁ(Φ)-PŁ problem, up to constant factors. This also applies to the function class $\mathcal{F}(L, \mu_1, \mu_2)$ since it is a subclass of PŁ(Φ)-PŁ ($\because$ Propositions 7 and 10).

Lemma 29. *Suppose $f(\boldsymbol{x};\boldsymbol{y})$ is an L-smooth function satisfying $\boldsymbol{y}$-side μ_2-PŁ condition and primal μ_1-PŁ condition (*i.e.*, PŁ(Φ)-PŁ). Suppose $\boldsymbol{z}^* = (\boldsymbol{x}^*;\boldsymbol{y}^*)$ is a global minimax point of f. Then, it satisfies*

$$\frac{\lambda\mu_1\mu_2^2}{2(\lambda\mu_1\mu_2+2L^2)}\left\|\boldsymbol{z}-\boldsymbol{z}^*\right\|^2 \le V_\lambda(\boldsymbol{z}) \le \frac{(\lambda+1)L^3}{\mu_2^2}\left\|\boldsymbol{z}-\boldsymbol{z}^*\right\|^2.$$

We remark that the second inequality also holds for general smooth nonconvex-PŁ problems.

Proof. Let $\kappa_1 = L/\mu_1$ and $\kappa_2 = L/\mu_2$ be condition numbers. By the conditions of f (smoothness and PŁ conditions), for any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\begin{aligned}
\frac{\mu_1}{2}\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 &\overset{\text{Prop. 10}}{\le} \Phi(\boldsymbol{x})-\Phi^* \overset{\text{Prop. 9}}{\le} \frac{L(\kappa_2+1)}{2}\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2,\\
\frac{\mu_2}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2 &\overset{\text{Ass. 4}}{\le} \Phi(\boldsymbol{x})-f(\boldsymbol{x};\boldsymbol{y}) \overset{\text{Ass. 1}}{\le} \frac{L}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2,
\end{aligned}$$

where $\boldsymbol{y}^*(\boldsymbol{x})$ is a projection of $\boldsymbol{y}$ to $\arg\max_{\boldsymbol{y}'} f(\boldsymbol{x};\boldsymbol{y}')$. In particular, $\boldsymbol{y}^*(\boldsymbol{x}^*) = \boldsymbol{y}^*$. Since $\boldsymbol{y}^*(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ and can differ from $\boldsymbol{y}^*$, we need to bound the term $\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2$ using $\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2$ and $\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2$. To upper-bound the term $\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2$, note that,

$$\begin{aligned}
\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2 &\le \left(\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x}^*)\right\| + \left\|\boldsymbol{y}^*(\boldsymbol{x})-\boldsymbol{y}^*(\boldsymbol{x}^*)\right\|\right)^2\\
&\le \left(\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\| + \kappa_2\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|\right)^2\\
&\le \left(1+\kappa_2^2\right)\left(\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2 + \left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2\right).
\end{aligned}$$

The first inequality holds by triangle inequality, the second inequality holds by Proposition 8, and the last inequality holds by Cauchy-Schwarz inequality.[11] To lower-bound in a similar way, note that for any constant $a > 0$,

$$\begin{aligned}
\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2 &\le \left(\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\| + \left\|\boldsymbol{y}^*(\boldsymbol{x})-\boldsymbol{y}^*(\boldsymbol{x}^*)\right\|\right)^2\\
&\le \left(\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\| + \frac{\kappa_2}{\sqrt{a}}\cdot\sqrt{a}\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|\right)^2\\
&\le \left(1+\frac{\kappa_2^2}{a}\right)\left(\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2 + a\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2\right).\\
\therefore \left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2 &\ge \frac{1}{1+\kappa_2^2/a}\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2 - a\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2.
\end{aligned}$$

Now we can prove the inequalities in the lemma. We first show the second one. Applying $\kappa_2 \ge 1$ multiple times,

$$\begin{aligned}
V_\lambda(\boldsymbol{x};\boldsymbol{y}) &= \lambda[\Phi(\boldsymbol{x})-\Phi^*] + [\Phi(\boldsymbol{x})-f(\boldsymbol{z})]\\
&\le \frac{\lambda L(\kappa_2+1)}{2}\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \frac{L}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2\\
&\le \left(\frac{\lambda L(\kappa_2+1)}{2} + \frac{L(1+\kappa_2^2)}{2}\right)\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \frac{L(1+\kappa_2^2)}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2\\
&\le (\lambda+1)L\kappa_2^2\left(\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2\right) = \frac{(\lambda+1)L^3}{\mu_2^2}\left\|\boldsymbol{z}-\boldsymbol{z}^*\right\|^2.
\end{aligned}$$

[11] $(ax+by)^2 \le (a^2+b^2)(x^2+y^2)$ for real numbers a, b, x, y.

To show the first inequality of the lemma, let $a = \frac{\lambda\mu_1}{2\mu_2}$.

$$\begin{aligned}
V_\lambda(\boldsymbol{x};\boldsymbol{y}) &\geq \frac{\lambda\mu_1}{2}\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \frac{\mu_2}{2}\left\|\boldsymbol{y}-\boldsymbol{y}^*(\boldsymbol{x})\right\|^2 \\
&\geq \left(\frac{\lambda\mu_1}{2} - \frac{\mu_2 a}{2}\right)\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \frac{\mu_2}{2(1+\kappa_2^2/a)}\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2 \\
&\geq \frac{\lambda\mu_1}{4}\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \frac{\lambda\mu_1}{4(a+\kappa_2^2)}\left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2 \\
&\geq \frac{\lambda\mu_1}{4(a+\kappa_2^2)}\left(\left\|\boldsymbol{x}-\boldsymbol{x}^*\right\|^2 + \left\|\boldsymbol{y}-\boldsymbol{y}^*\right\|^2\right) = \frac{\lambda\mu_1\mu_2^2}{2(\lambda\mu_1\mu_2+2L^2)}\left\|\boldsymbol{z}-\boldsymbol{z}^*\right\|^2.
\end{aligned}$$

This concludes the proof. □

The equivalence of iteration complexities for achieving $\left\|\boldsymbol{z}_K - \boldsymbol{z}^*\right\|^2 \leq \varepsilon^2$ or $V_\lambda(\boldsymbol{z}_K) \leq \varepsilon^2$ is quite straightforward from this lemma, as long as the convergence speed is exponential. For example, suppose we have a upper convergence bound $\left\|\boldsymbol{z}_K - \boldsymbol{z}^*\right\|^2 \leq a\exp(-K/r)$ for some constants $a, r > 0$. This implies a upper iteration complexity bound $K = \mathcal{O}(r\log(1/\varepsilon))$ sufficient to achieve $\left\|\boldsymbol{z}_K - \boldsymbol{z}^*\right\|^2 \leq \varepsilon^2$. Then by Lemma 29, we also have $V_\lambda(\boldsymbol{z}_K)^2 \leq a'\exp(-K/r)$ where $a' = a(\lambda+1)L^3/\mu_2^2$ is also a constant. This implies a lower iteration complexity bound $K = \mathcal{O}(r\log(1/\varepsilon))$ as well, sufficient to achieve $V_\lambda(\boldsymbol{z}_K)^2 \leq \varepsilon^2$. The other way of complexity translation operates with a similar logic.

F OMITTED COMPARISON WITH RELATED WORKS

F.1 COMPARISON WITH XIE ET AL. (2021)

To specialize Xie et al. (2021, Theorem 3) to the single-machine setup and discuss their results in terms of our notation, we need to replace their symbols

$$(T, S, K, \sigma_1^2, \sigma_2^2, G_1^2, G_2^2, L_{12}, L_f, \mu, L_\Phi, \mathcal{L}_0, \eta_t, \gamma_t)$$

with the following symbols from our notation

$$(K, 1, n, 0, 0, B, B, L, L, \mu_2, L(\kappa_2+1), V_\lambda(\boldsymbol{z}_0^1), \alpha, \beta),$$

and also put $A = 0$ (their analysis only applies uniformly bounded component variance per machine). Then we can *naively* translate the bound of Xie et al. (2021, Theorem 3) to our language as

$$\min_{k\in[K]} \mathbb{E}\left[\|\Phi(\boldsymbol{x})\|^2\right] \overset{(?)}{\leq} \mathcal{O}\left(\frac{\kappa_2 L V_\lambda(\boldsymbol{z}_0^1)}{K} + \kappa_2^2\left(\frac{L^2 B V_\lambda(\boldsymbol{z}_0^1)^2}{K^2}\right)^{1/3}\right).$$

To the best of our knowledge, however, we believe there may be a mistake in the proof of Xie et al. (2021, Appendix C.4). From the inequalities in the last page of their paper, we notice that the term $\frac{40L_{12}^2\mathcal{L}_0}{\mu^2\gamma KT}$ might be missing in a step, where γ is chosen to be the minimum of several terms including $\frac{1}{87L_f K}$. Thus, as far as we can tell, it seems inevitable that this omitted term would lead to an additional term $\frac{3480L_f L_{12}^2\mathcal{L}_0}{\mu^2 T}$ in the final bound. By combining this to their bound and re-translating it, we eventually have

$$\min_{k\in[K]} \mathbb{E}\left[\|\Phi(\boldsymbol{x})\|^2\right] \leq \mathcal{O}\left(\frac{\kappa_2^2 L V_\lambda(\boldsymbol{z}_0^1)}{K} + \kappa_2^2\left(\frac{L^2 B V_\lambda(\boldsymbol{z}_0^1)^2}{K^2}\right)^{1/3}\right),$$

since their L_{12}^2/μ^2 translates to our κ_2^2. Therefore, their result actually shows the same dependency on condition number κ_2 as our Theorem 1. Nevertheless, comparing the terms related to the component-wise variance B, ours is better. In the second term in the bound above does not shrink even when the number of iterations (per machine & per communication) grows. In our case (Theorem 1), however, the dominant term (in K) can be briefly written as $\mathcal{O}\left(\left(\frac{B}{nK^2}\right)^{1/3}\right)$ which can diminish with large n, *i.e.*, the number of iterations per epoch.

G DESCRIPTION ON EXPERIMENTS

In this appendix, we provide a more detailed illustration of our numerical evaluations introduced in Section 6. Recall that the objective function f and its component functions f_i are given in Equation (5) as

$$f(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y},$$
$$f_i(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{B}_i\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{C}_i\boldsymbol{y} + \boldsymbol{u}_i^\top \boldsymbol{x} - \boldsymbol{v}_i^\top \boldsymbol{y}.$$

We choose the same dimensions for the variables $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$; *i.e.*, we set $d_x = d_y = d$.

G.1 PARAMETER CHOICES

To sample the matrix $\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{C}_i \in \mathbb{R}^d$ satisfying that $\mu_C \boldsymbol{I}_d \preceq \boldsymbol{C}$ and $\|\boldsymbol{C}_i\|_2 \le L_C$, we first randomly generate an orthogonal matrix $\boldsymbol{Q}_C \in \mathbb{R}^{d\times d}$ (*i.e.*, $\boldsymbol{Q}_C\boldsymbol{Q}_C^\top = \boldsymbol{I}_d$), by taking advantage of the QR-decomposition a random matrix. Then, we generate the eigenvalues of $\boldsymbol{C}_i$'s as follows. We sample the entries of n vectors $\boldsymbol{\lambda}_i^C \in \mathbb{R}^d$ $(i = 1, \cdots, n)$ uniformly from the interval $[\mu_C, L_C]$. We add some level of perturbations to some entries of each $\boldsymbol{\lambda}_i^C$; we replace some entries to the numbers in an interval $[-L_C, \mu_C]$, keeping the entries of the vector $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}_i^C$ in the interval $[\mu_C, L_C]$. Finally, we define $\boldsymbol{C}_i = \boldsymbol{Q}_C\boldsymbol{\Lambda}_i^C\boldsymbol{Q}_C^\top$ where $\boldsymbol{\Lambda}_i^C = \mathrm{diag}(\boldsymbol{\lambda}_i^C)$. Because of the perturbation step, some $\boldsymbol{C}_i$'s are not positive definite, thereby some components f_i's become non-(strongly-)concave in $\boldsymbol{y}$.

Next, we sample the matrix $\boldsymbol{B}_i$'s. There are no requirements for $\boldsymbol{B}$ but $\|\boldsymbol{B}_i\|_2 \le L_B$; $\boldsymbol{B}_i$'s are even not necessarily symmetric. Thus, we first generate the orthogonal matrices $\boldsymbol{U}_i^B$ and $\boldsymbol{V}_i^B$ by taking advantage of the singular value decomposition of random matrices. Then, we generate the singular values of $\boldsymbol{B}_i$'s by sampling the entries of n vectors $\boldsymbol{\sigma}_i^B$ uniformly from the interval $[0, L_C]$. After that, we define $\boldsymbol{B}_i = \boldsymbol{U}_i^B\boldsymbol{\Sigma}_i^B\boldsymbol{V}_i^B$ where $\boldsymbol{\Sigma}_i^C = \mathrm{diag}(\boldsymbol{\sigma}_i^C)$. We typically want to take a larger L_B than L_C to strengthen the interaction term $\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{y}$.

Recall that the primal function Φ associated with f is explicitly written as

$$\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathbb{R}^d} f(\boldsymbol{x};\boldsymbol{y}) = \tfrac{1}{2}\boldsymbol{x}^\top\left(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top\right)\boldsymbol{x} := \tfrac{1}{2}\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}. \tag{28}$$

Before generating the matrices $\boldsymbol{A}_i$'s, we first generate $\boldsymbol{M}_i$'s $(i = 1, \cdots, n)$ satisfying that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{M}_i = \boldsymbol{M}$ and that the nonzero eigenvalues of positive *semi*definite $\boldsymbol{M}$ are in the interval $[\mu_M, L_M]$. The process of sampling $\boldsymbol{M}_i$'s is almost identical to how to sample $\boldsymbol{C}_i$'s. One notable difference is that $\boldsymbol{M}_i$'s and $\boldsymbol{M}$ are forced to have $r(< d)$ zero eigenvalues: this makes $\boldsymbol{M}$ a positive semidefinite (but not strictly positive definite) matrix of rank $d - r$. Moreover, we get the μ_M-PŁ(Φ) condition in $\boldsymbol{x}$ as follows:

Proposition 30. *Consider a positive semidefinite matrix $\boldsymbol{M} \in \mathbb{R}^d$. If the smallest nonzero eigenvalue of $\boldsymbol{M}$ is μ, then $\Phi(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}$ is μ-PŁ in $\boldsymbol{x}$.*

Proof. Apply the eigendecomposition of $\boldsymbol{M}$: $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$. Let $\overline{\boldsymbol{M}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top$. Then, we have $\Phi(\boldsymbol{x}) = \frac{1}{2}\left\|\overline{\boldsymbol{M}}\boldsymbol{x}\right\|^2$. Note that $\frac{1}{2}\|\boldsymbol{x}\|^2$ is 1-strongly convex. Also, the minimum nonzero singular value of $\overline{\boldsymbol{M}}$ is $\sqrt{\mu}$ ($\because \boldsymbol{M} = \overline{\boldsymbol{M}}^\top\overline{\boldsymbol{M}}$). Therefore, by the proof of Proposition 12, $\Phi(\boldsymbol{x})$ is a μ-PŁ function of $\boldsymbol{x}$. Lastly, we note that $\Phi(\boldsymbol{x})$ is not strongly-convex in general, especially when $\boldsymbol{M}$ is a rank-deficient matrix. □

Typically, the spectral norm $\|\boldsymbol{M}\|_2$ is known to be bounded above by $\|\boldsymbol{A}\|_2 + L_B^2/\mu_C$ in *worst-case* (Nouiehed et al., 2019; Li et al., 2022). However, since we sample $\boldsymbol{M}$ without knowing the exact form of $\boldsymbol{A}_i$'s but we want to control the spectral norm $\|\boldsymbol{A}_i\|_2$ not too large (for smoothness of f_i), we decided to choose rather smaller L_M: simply, we chose $L_M = L_B$.

Now we let $\boldsymbol{A}_i = \boldsymbol{M}_i - \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top$ and $\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i$ to satisfy Equation (28). We emphasize that $\boldsymbol{A}$ may have negative eigenvalues; the objective is nonconvex in $\boldsymbol{x}$ in general. We have checked this is true across the experimental settings. Also, we let $L := \max\{\|L_A\|_2, L_B, L_C\}$ for further

parameter selection. (In fact, because of our choice of parameter values, L was always equal to L_B in our experiments.)

Furthermore, we generate the vectors $\boldsymbol{u}_i$'s and $\boldsymbol{v}_i$'s satisfying $\sum_{i=1}^n \boldsymbol{u}_i = \mathbf{0} = \sum_{i=1}^n \boldsymbol{v}_i$. The entries of these vectors are uniformly sampled from the interval $[-\Delta, \Delta]$, and then we centered the average of entries to be zero.

To verify our theory, we choose the step-sizes of the form $\beta = c_1 \cdot \frac{1}{nL}$ and $\alpha = c_0 \cdot \frac{\beta}{\kappa_2^2}$ for some constants c_0 and c_1.

Lastly, we specify the exact values of some parameters described above: $n = 100$, $d = 25$, $c_0 = c_1 = 1$, $\mu_M = \mu_C$, and $L_C = 1 < L_M = L_B$. Besides, we used mini-batches of size 1. In the following subsections, we investigate the effects of change in (i) Δ, determining the discrepancy between components, and (ii) the condition number κ_2, determined by L_B and μ_C.

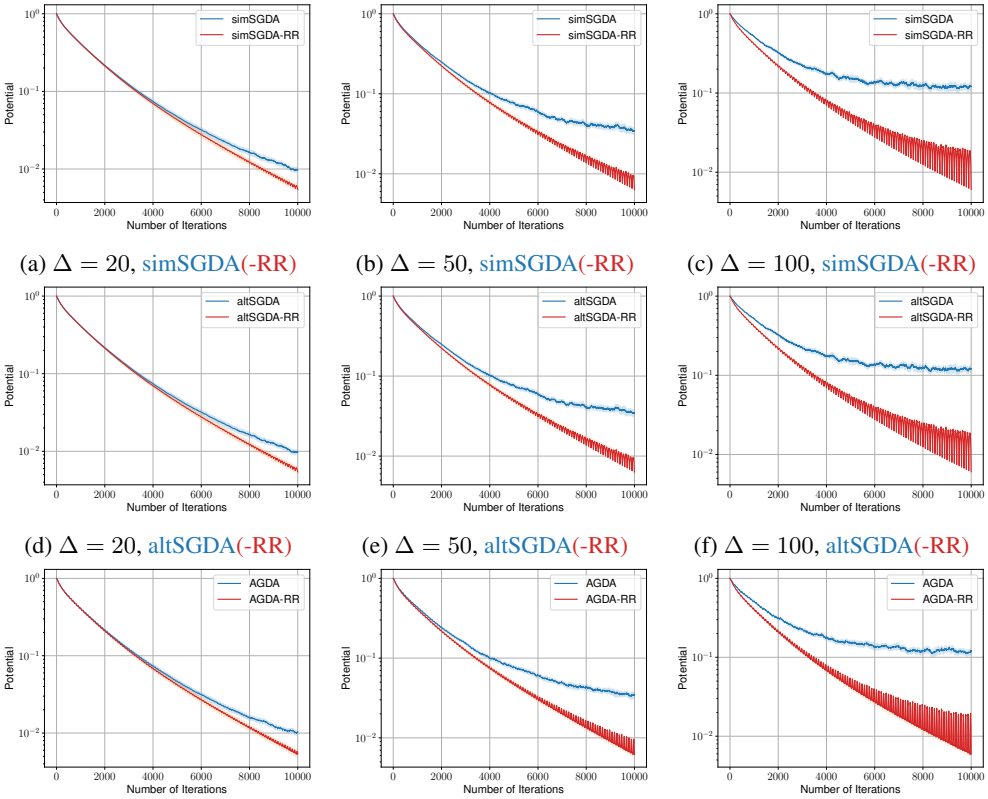
G.2 Comparison: the effect of component discrepancy

Notice that the discrepancy between components gets larger as Δ grows. Technically, one can check that the gradient variance (that we controlled in Assumption 2) is proportional to the norms of the vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$. Moreover, we have already discussed that SGDA with random reshuffling can converge much faster than its with-replacement counterpart especially when the gradient variance is large.

Now, we present the results of numerical experiments by varying the values of Δ to 20, 50, and 100, while fixing $L_B = 2$, $\mu_C = 0.5$, and other experiment parameters. As shown in Figure 2, we can observe that the difference between the random-reshuffling algorithm and the uniform-sampling algorithm gets larger as Δ increases.

G.3 Comparison: the effect of condition number

Here, we present the results of numerical experiments by varying the values of κ_2 to 2, 4, and 8, while fixing $\Delta = 50$ and other experiment parameters. To this end, we applied the parameter settings for L_B and μ_C as $(L_B, \mu_C) = (1, 0.5), (2, 0.5), (2, 0.25)$. The results are shown in Figure 3. We observe that more epochs are required for convergence when κ_2 increases, regardless of the type of algorithm. One may think that the performance gap between RR-based/non-RR-based algorithms is small when κ_2 is huge. However, when we run the algorithm for an extended number of epochs, we observe a significant gap in convergence speeds.

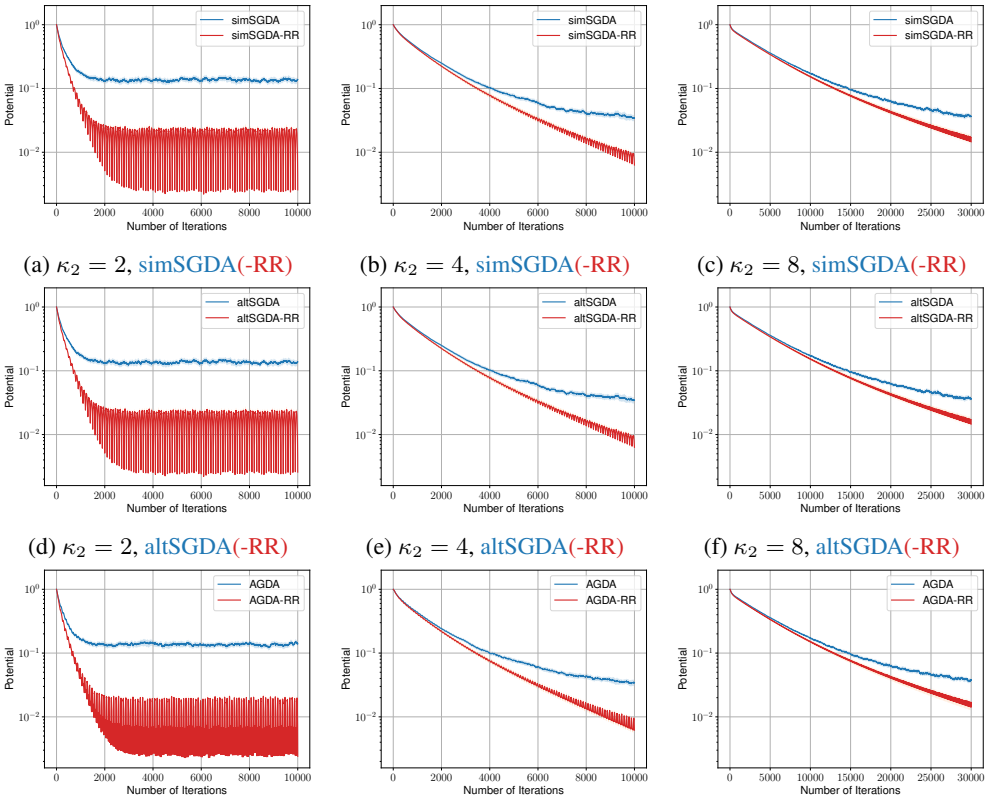


(a) $\Delta = 20$, simSGDA(-RR) (b) $\Delta = 50$, simSGDA(-RR) (c) $\Delta = 100$, simSGDA(-RR)

(d) $\Delta = 20$, altSGDA(-RR) (e) $\Delta = 50$, altSGDA(-RR) (f) $\Delta = 100$, altSGDA(-RR)

(g) $\Delta = 20$, AGDA(-RR) (h) $\Delta = 50$, AGDA(-RR) (i) $\Delta = 100$, AGDA(-RR)

Figure 2: Comparisons by changing the value of $\Delta \in \{20, 50, 100\}$. The solid lines are the average across 50 different quadratic games and 95% confidence intervals (± 1.96 std) are shown as the shaded region. The values are normalized by dividing them by the initial value. Note that the vertical axes are on a logarithmic scale.



(a) $\kappa_2 = 2$, simSGDA(-RR) (b) $\kappa_2 = 4$, simSGDA(-RR) (c) $\kappa_2 = 8$, simSGDA(-RR)

(d) $\kappa_2 = 2$, altSGDA(-RR) (e) $\kappa_2 = 4$, altSGDA(-RR) (f) $\kappa_2 = 8$, altSGDA(-RR)

(g) $\kappa_2 = 2$, AGDA(-RR) (h) $\kappa_2 = 4$, AGDA(-RR) (i) $\kappa_2 = 8$, AGDA(-RR)

Figure 3: Comparisons by changing the value of $\kappa_2 = L/\mu_C \in \{2, 4, 8\}$. The solid lines are the average across 50 different quadratic games and 95% confidence intervals (± 1.96 std) are shown as the shaded region. The values are normalized by dividing them by the initial value. Note that the vertical axes are on a logarithmic scale. In addition, we run 300 epochs for $\kappa_2 = 8$ (see the rightmost column), whereas we run 100 epochs for the other $\kappa_2 \in \{2, 4\}$ (see the leftmost & middle columns).