Sampling without Replacement Leads to Faster Rates in Finite-Sum Minimax Optimization

Aniket Das *
Indian Institute of Technology Kanpur
aniketd@iitk.ac.in

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
bs@tuebingen.mpg.de

Michael Muehlebach
Max Planck Institute for Intelligent Systems
michaelm@tuebingen.mpg.de

Abstract

We analyze the convergence rates of stochastic gradient algorithms for smooth finite-sum minimax optimization and show that, for many such algorithms, sampling the data points *without replacement* leads to faster convergence compared to sampling with replacement. For the smooth and strongly convex-strongly concave setting, we consider gradient descent ascent and the proximal point method, and present a unified analysis of two popular without-replacement sampling strategies, namely *Random Reshuffling* (RR), which shuffles the data every epoch, and *Single Shuffling* or *Shuffle Once* (SO), which shuffles only at the beginning. We obtain tight convergence rates for RR and SO and demonstrate that these strategies lead to faster convergence than uniform sampling. Moving beyond convexity, we obtain similar results for smooth nonconvex-nonconcave objectives satisfying a two-sided Polyak-Łojasiewicz inequality. Finally, we demonstrate that our techniques are general enough to analyze the effect of *data-ordering attacks*, where an adversary manipulates the order in which data points are supplied to the optimizer. Our analysis also recovers tight rates for the *incremental gradient* method, where the data points are not shuffled at all.

1 Introduction

The approximate solution of large-scale optimization problems using first-order stochastic gradient methods constitutes one of the foundations of classical machine learning. However, emerging problems in machine learning go beyond pattern recognition and involve real-world decision making, where learning algorithms interact with unknown or even adversarial environments or are deployed in multi-agent settings. Decision making in such environments often involves solving a *minimax optimization* problem of the form $\min_{\mathbf{x}} \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$, whose analysis has been a focus of research in mathematics, economics, and theoretical computer science [38, 13, 10]. Recent examples of its applications in machine learning include adversarial learning [32, 52, 6], reinforcement learning [30, 58, 11, 42], imitation learning [14, 8, 24], and generative adversarial networks [18, 2]. In most of these applications, the objective $F(\mathbf{x}, \mathbf{y})$ has a finite-sum structure, i.e., $F(\mathbf{x}, \mathbf{y}) = 1/n \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y})$ where n denotes the size of the dataset and each component function f_i denotes the objective associated with the ith data point. The resultant problem is known as *finite-sum minimax optimization*:

---

*Currently at Google Research. Contact: ketd@google.com

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

$$\min_{\mathbf{x}\in\mathbb{R}^{d_{\mathbf{x}}}} \max_{\mathbf{y}\in\mathbb{R}^{d_{\mathbf{y}}}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}). \tag{1}$$

When n is large and the f_i's are differentiable (which holds for most applications), approximate solutions to (1) are computed using stochastic gradient algorithms. These algorithms typically sample an index $i \in [n]$ at each iteration as per some specified sampling routine, and use the gradients of f_i to compute the next iterate. Among these methods, perhaps the simplest and most commonly used algorithm is Stochastic Gradient Descent Ascent (SGDA), a natural extension of Stochastic Gradient Descent (SGD) for minimax optimization.

Similar to the stochastic optimization literature, analysis of stochastic minimax optimization often assumes that, at every iteration, the index i is sampled *uniformly with replacement*. Analysis of the resulting algorithm closely parallels the analysis of SGD, and relies on the fact that uniform sampling leads to unbiased gradient estimates. Recent works [20, 31] have also extended this paradigm to i.i.d uniform sampling of mini-batches. While uniform sampling assumptions simplify theoretical analysis, practical implementations of these algorithms often deviate from this paradigm, and instead incorporate various heuristics, which are often empirically found to improve runtime. A common and notable heuristic is to replace uniform sampling by procedures that perform multiple passes over the entire dataset, and in each such pass (called an epoch), sample the data points *without replacement*. Thus, each data point is sampled exactly once in every epoch. These procedures are generally implemented using one of the following approaches:

Random Reshuffling (RR): Uniformly sample a random permutation of $[n]$ at the start of every epoch, and process the data points within that epoch as per the order specified by the permutation.

Single Shuffling or **Shuffle Once** (SO): Uniformly sample a random permutation at the beginning and reuse it across all epochs to order the data points.

Incremental Gradient (IG): Do not permute the data points at all and follow a fixed deterministic data ordering for every epoch.

Sampling without replacement is ubiquitous in both stochastic minimization [7, 48, 4] and stochastic minimax optimization [18, 2] as it often exhibits faster runtime than uniform sampling. However, these empirical benefits come at the cost of limited theoretical understanding, due to the absence of provably unbiased gradient estimates.

It is well known in the *optimization* literature that SGD with replacement has a tight rate of $O(1/nK)$ for smooth and strongly convex minimization [44, 25], where n is the number of component functions and K denotes the number of epochs. On the contrary, recent works on SGD without replacement for smooth and strongly convex minimization [1, 36, 34, 39] show that both RR and SO achieve a non-asymptotic rate of $\tilde{\mathcal{O}}(1/nK^2)$, once the number of epochs K is larger than a certain threshold K_0 (usually polynomial in the condition number), and thereby converge faster than SGD with replacement. These rates match the lower bound of $\Omega(1/nK^2)$ for RR and SO established in prior works [43, 48], modulo logarithmic factors. For RR, prior works have also established a similar $\tilde{\mathcal{O}}(1/nK^2)$ rate for nonconvex objectives satisfying the Polyak-Łojasiewicz (PŁ) inequality [1, 34]. While the asymptotic behavior of IG has been known to the community for a long time in both smooth and non-smooth settings [5, 37], non-asymptotic $\tilde{\mathcal{O}}(1/K^2)$ convergence rates have been established quite recently [34, 39, 22], and are complemented by a matching $\Omega(1/K^2)$ lower bound [48].

1.1 Contributions

Although the empirical benefits of sampling without replacement have been substantiated for minimization, analysis of these methods for *minimax optimization* have received much less attention, despite being widely prevalent in many applications. Our work aims to fill this gap by analyzing these methods for minimax optimization. To this end, our main contributions are as follows:

Unified analysis of RR and SO for smooth strongly convex-strongly concave problems: We analyze RR and SO in conjunction with simultaneous *Gradient Descent Ascent* (GDA), calling the resulting algorithms GDA-RR and GDA-SO, respectively. Assuming the components f_i are smooth and F is strongly convex-strongly concave, we present a unified analysis of GDA-RR/SO and establish a convergence rate of $\tilde{O}(\exp(-K/5\kappa^2) + 1/nK^2)$ for both (where κ is the condition

2

number). Comparing with lower bounds, we show that our rates are *nearly tight*, i.e., they differ from the lower bound only by an exponentially decaying term. Moreover, when $K \geq 10\kappa^2 \log(n^{1/2}K)$, the convergence rate matches the lower bounds for GDA-RR/SO, modulo logarithmic factors, and also converges provably faster than SGDA with replacement. Under the same setting, we obtain similar guarantees for the RR and SO variants of the *Proximal Point Method* (PPM), named PPM-RR and PPM-SO respectively. Our analysis for both GDA-RR/SO and PPM-RR/SO is general enough to cover smooth strongly monotone finite-sum variational inequalities, which covers minimization, minimax optimization, and multiplayer games.

RR for smooth two-sided PŁ objectives: We consider a class of nonconvex-nonconcave problems where the objective F satisfies a *two-sided Polyak-Łojasiewicz* inequality. For such problems, we propose an algorithm that combines RR with two-timescale *Alternating Gradient Descent Ascent* (AGDA), which we call AGDA-RR. We show that AGDA-RR has a nearly tight convergence rate of $\tilde{O}(\exp(-K/365\kappa^3) + 1/nK^2)$ when the gradient variance is uniformly bounded. When $K \geq 730\kappa^3 \log(n^{1/2}K)$, this rate matches the lower bound (modulo logarithmic factors) and improves on the best known rates of with-replacement algorithms for this class of problems.

Minimax optimization under data ordering attacks: Our techniques for analyzing RR/SO generalize to the analysis of finite-sum minimax optimization under *data ordering attacks* [51]. These attacks target the inherent randomness assumptions of stochastic gradient algorithms, significantly increasing training time and reducing model quality, only by manipulating the order in which the algorithm receives data points, without performing any data contamination. To model these attacks, we propose the *Adversarial Shuffling* (AS) setup, where the data points are shuffled every epoch by a computationally unrestricted adversary. In this setup, we show that GDA and PPM (now called GDA-AS and PPM-AS) have a convergence rate of $\tilde{O}(\exp(-K/5\kappa^2) + 1/K^2)$ for smooth strongly convex-strongly concave objectives, and AGDA (now called AGDA-AS) has a rate of $\tilde{O}(\exp(-K/365\kappa^3) + 1/K^2)$ for two-sided PŁ objectives. We note that, compared to RR and SO, the convergence rate worsens by a factor of $1/n$ for large enough K. When n is large (true for most applications), this slowdown significantly impacts convergence and thus, theoretically justifies the empirical observations in prior work [51]. We also establish that our analysis in the AS regime also applies to the Incremental Gradient (IG) variants of these algorithms (namely GDA-IG, PPM-IG, and AGDA-IG), and use this to show that our obtained rates for GDA-RR and AGDA-RR are nearly tight.

To the best of our knowledge, our work is the first to: 1) analyze RR, SO, and IG for strongly monotone unconstrained variational inequalities, 2) analyze RR and IG for a class of nonconvex-nonconcave minimax problems, 3) provably demonstrate the advantages of sampling without replacement for both these settings and justify its empirical benefits in a wide variety of problems ranging from minimization, minimax optimization to smooth multiplayer games, 4) analyze sampling without replacement under data-ordering attacks. Furthermore, unlike prior works on sampling without replacement for minimax optimization [55, 33], which are restricted to random reshuffling and require the component functions to be convex-concave, Lipschitz, and smooth, our analysis does not impose any restrictions on the components f_i other than smoothness, allowing them to be arbitrary nonconvex-nonconcave functions.

2 Notation and Preliminaries

We work with Euclidean spaces $(\mathbb{R}^d, \langle ., . \rangle)$ equipped with the standard inner product $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle$ and the induced norm $|\mathbf{x}|$. For any $\mathbf{x} \in \mathbb{R}^{d_{\mathbf{x}}}$ and $\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}$, we denote $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ where $d = d_{\mathbf{x}} + d_{\mathbf{y}}$. Moreover, for any $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1) \in \mathbb{R}^d$ and $\mathbf{z}_2 = (\mathbf{x}_2, \mathbf{y}_2) \in \mathbb{R}^d$, $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle + \langle \mathbf{y}_1, \mathbf{y}_2 \rangle$ and $|\mathbf{z}_1|^2 = |\mathbf{x}_1|^2 + |\mathbf{y}_1|^2$. Whenever $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is clear from the context, we write $f(\mathbf{x}, \mathbf{y})$ as $f(\mathbf{z})$. We use $\mathbb{S}_n$ to denote the set of all permutations of $[n] = \{1, \ldots, n\}$. For any matrix $\mathbf{A}$, its operator norm is denoted by $|\mathbf{A}| = \sup_{|\mathbf{x}|=1} |\mathbf{A}\mathbf{x}|$. We use the O notation to characterize the dependence of our convergence rates on n and K, suppressing numerical and problem-specific constants such as κ, μ, σ, etc. Additionally, we use the $\tilde{O}$ notation to suppress logarithmic factors of n and K.

Our work studies finite-sum minimax optimization (1). Solutions to (1) are known as *global minimax points* of $F = 1/n \sum_{i=1}^{n} f_i$, which *we assume to always exist*. We also assume that the components f_i are continuously differentiable, and hence, the same applies to F. This allows us to define the

3

gradient operators $\omega_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ *and* $\nu : \mathbb{R}^d \rightarrow \mathbb{R}^d$ *as follows:*

$$\omega_i(\mathbf{x}, \mathbf{y}) = [\nabla_{\mathbf{x}} f_i(\mathbf{x}, \mathbf{y}), -\nabla_{\mathbf{y}} f_i(\mathbf{x}, \mathbf{y})], \quad \nu(\mathbf{x}, \mathbf{y}) = 1/n \sum_{i=1}^{n} \omega_i(\mathbf{x}, \mathbf{y}).$$

We also impose the following smoothness assumption on the components f_i.

Assumption 1 (Component Smoothness). *The component functions f_i are l-smooth, i.e., each gradient operator ω_i is l-Lipschitz*

$$|\omega_i(\mathbf{z}_2) - \omega_i(\mathbf{z}_1)| \leq l \,|\mathbf{z}_2 - \mathbf{z}_1|.$$

Consequently, the operator ν is also l-Lipschitz, i.e., F is l-smooth.

3   Analysis for Strongly Convex-Strongly Concave Objectives

In this section, we analyze two very popular without-replacement algorithms for finite-sum minimax optimization, Gradient Descent Ascent (GDA) without replacement and Proximal Point Method (PPM) without replacement. For each of these, we present a unified analysis of the Random Reshuffling (RR) and Shuffle Once (SO) variants (called GDA-RR/SO and PPM-RR/SO respectively). For a fixed $K > 0$, GDA-RR/SO approximately solves (1) by iterating over the entire dataset for K epochs, and within each epoch, uses the operators ω_i to perform the following iterative update:

$$\mathbf{z}_i^k \leftarrow \mathbf{z}_{i-1}^k - \alpha \omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k) \,\forall i \in [n], \tag{2}$$

where τ_k is a uniformly sampled random permutation of $[n]$ and $0 < \alpha < 1/l$ is a constant step-size. GDA-RR resamples τ_k at the start of every epoch, whereas GDA-SO samples it only once in the beginning. The details of both algorithms are presented in Algorithm 1. The Proximal Point Method without replacement is a closely related algorithm which, instead of performing gradient descent-style updates within an epoch, solves the following implicit update equation for $\mathbf{z}_i^k$:

$$\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha \omega_{\tau_k(i)}(\mathbf{z}_i^k) \,\forall i \in [n]. \tag{3}$$

As before, τ_k is resampled at every epoch for PPM-RR, and sampled once and fixed for all epochs for PPM-SO. We present the details in Algorithm 2. The l-smoothness of ω_i along with the choice of $\alpha < 1/l$ ensures that $\mathbf{z}_i^k$ is uniquely defined, since it is a fixed point of the contraction mapping $\zeta(\mathbf{z}) = \mathbf{z}_{i-1}^k - \alpha \omega_{\tau_k(i)}(\mathbf{z})$. This method is actually a generalization of the (stochastic) proximal point method for minimization problems, and is popular for problems where (3) can be solved easily or in closed form. We refer the readers to Rockafellar [47], Patrascu and Necoara [41] for a review of this method and its connections to the original proximal point method for minimization.

3.1   Setting

We analyze GDA-RR/SO and PPM-RR/SO for smooth finite-sum strongly convex-strongly concave (or SC-SC) objectives. This allows us to formulate the minimax optimization problem for F as a *root finding problem* for the gradient operator ν, as described below.

Assumption 2 (Strong Convexity-Strong Concavity). *The objective F is μ strongly convex-strongly concave (or SC-SC), i.e., $F(., \mathbf{y})$ is μ-strongly convex for any $\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}$ and $-F(\mathbf{x}, .)$ is μ-strongly convex for any $\mathbf{x} \in \mathbb{R}^{d_{\mathbf{x}}}$.*

Assumption 2 has the following consequences for the gradient operator ν:

Lemma 1. *Let F satisfy Assumptions 1 and 2. Then, the gradient operator ν is μ-strongly monotone:*

$$\langle \nu(\mathbf{z}_1) - \nu(\mathbf{z}_2), \mathbf{z}_1 - \mathbf{z}_2 \rangle \geq \mu \,|\mathbf{z}_1 - \mathbf{z}_2|^2 \,\forall \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d.$$

Furthermore, (1) admits a unique solution $\mathbf{z}^*$, which is also the unique solution of $\nu(\mathbf{z}^*) = 0$.

Lemma 1 allows us to recast (1) for SC-SC objectives as the following *root finding problem*:

$$\text{Find } \mathbf{z} \in \mathbb{R}^d \text{ such that } \nu(\mathbf{z}) = 1/n \sum_{i=1}^{n} \omega_i(\mathbf{z}) = 0. \tag{4}$$

4

Algorithm 1: GDA-RR/SO/AS	**Algorithm 2:** PPM-RR/SO/AS
Input : Number of epochs K, step-size $\alpha > 0$, and initialization $\mathbf{z}_0$	**Input** : Number of epochs K, step-size $\alpha > 0$, and initialization $\mathbf{z}_0$
Initialize $\mathbf{z}_0^1 \leftarrow \mathbf{z}_0$	Initialize $\mathbf{z}_0^1 \leftarrow \mathbf{z}_0$
SO: Sample $\tau \sim \text{Uniform}(\mathbb{S}_n)$	SO: Sample $\tau \sim \text{Uniform}(\mathbb{S}_n)$
for $k \in [K]$ **do**	**for** $k \in [K]$ **do**
  RR: Sample $\tau_k \sim \text{Uniform}(\mathbb{S}_n)$	 RR: Sample $\tau_k \sim \text{Uniform}(\mathbb{S}_n)$
  SO: $\tau_k \leftarrow \tau$	 SO: $\tau_k \leftarrow \tau$
  AS: Adversary chooses $\tau_k \in \mathbb{S}_n$	 AS: Adversary chooses $\tau_k \in \mathbb{S}_n$
  for $i \in [n]$ **do**	 **for** $i \in [n]$ **do**
    $\mathbf{z}_i^k \leftarrow \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k)$	 Solve the implicit update for $\mathbf{z}_i^k$ where, $\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_i^k)$
  end	 **end**
  $\mathbf{z}_0^{k+1} \leftarrow \mathbf{z}_n^k$	 $\mathbf{z}_0^{k+1} \leftarrow \mathbf{z}_n^k$
end	**end**

Figure 1: GDA-RR/SO/AS and PPM-RR/SO/AS for solving (4). Violet lines denote steps that are only performed for RR, Olive lines denote the same for SO and Magenta for AS.

Problem (4) is more general than SC-SC minimax optimization, and is a special case of strongly monotone variational inequalities [13] without constraints. Notably, (4) includes the Nash Equilibrium problem for unconstrained *multiplayer* games with smooth strongly convex objectives [50] and is sometimes called a *finite-sum unconstrained variational inequality* in the literature [31]. We also highlight that smooth strongly convex *optimization* is a special case of (4). However, unlike the optimization setting, ν is no longer restricted to be the gradient of a strongly convex function, which, as we shall see, has important consequences for the attainable convergence rates of our algorithms.

3.2 Analysis of RR/SO

We now state the expected last-iterate convergence guarantees for Algorithms 1 and 2 for solving (4), where the expecation is taken over the stochasticity of the sampled permutation(s).

Theorem 1 (Convergence of GDA-RR/SO and PPM-RR/SO). *Consider Problem* (4) *for the μ-strongly monotone operator $\nu(\mathbf{z}) = {}^1\!/n \sum_{i=1}^n \omega_i(\mathbf{z})$ where each ω_i is l-Lipschitz, but not necessarily monotone. Let $\mathbf{z}^*$ denote the unique root of ν. Then, there exists a step-size $\alpha \leq {}^\mu\!/5nl^2$ for which both GDA-RR/SO and PPM-RR/SO satisfy the following for any $K \geq 1$:*

$$\mathbb{E}[|\mathbf{z}_0^{K+1} - \mathbf{z}^*|^2] \leq 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 8\kappa^2\sigma_*^2\log^3(|\nu(\mathbf{z}_0)|n^{1/2}K/\mu)}{\mu^2 nK^2} = \tilde{O}(e^{-K/5\kappa^2} + {}^1\!/nK^2),$$

where $\kappa = {}^l\!/\mu$ is the condition number and $\sigma_*^2 = {}^1\!/n \sum_{i=1}^n |\omega_i(\mathbf{z}^*)|^2$ is the gradient variance at $\mathbf{z}^*$.

Proof. We present an outline for GDA-RR/SO and defer the full proof to Appendix C.2 (for GDA-RR/SO) and Appendix D.2 (for PPM-RR/SO). Furthermore, we recall that the updates of GDA-RR/SO are given by $\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k)$.

We begin with the following key insight from earlier works on sampling without replacement for minimization [23, 1, 37, 22]: for small enough step-sizes, the *epoch iterates* $\mathbf{z}_0^k$ of GD without replacement approximately follow the trajectory of full-batch gradient descent. To this end, we derive the following *epoch-level* update rule for GDA-RR/SO by linearizing $\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k)$ around $\mathbf{z}^*$:

$$\mathbf{z}_0^{k+1} - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^k - \mathbf{z}^*) + \alpha^2 \mathbf{r}_k, \tag{5}$$

where $|\mathbf{H}_k| \leq 1 - n\alpha\mu/2$ and $\mathbf{r}_k = \sum_{i=1}^{n-1} \mathbf{A}_{\tau_k(i)} \sum_{j=1}^i \omega_{\tau_k(j)}(\mathbf{z}^*)$ with $|\mathbf{A}_{\tau_k(i)}| \leq le^{1/5}$. The term $\mathbf{r}_k$ encapsulates the noise of the stochastic gradient updates accumulated over an entire epoch. To ensure convergence, we control the influence of the noise term $\mathbf{r}_k$ by using standard properties of without-replacement sample means to show that $\mathbb{E}[|\mathbf{r}_k|^2] \leq l^2 n^3 \sigma_*^2/4$ for both RR and SO. We then complete the proof by unrolling (5) for K epochs, substituting the upper bounds for $|\mathbf{H}_k|$ and $\mathbb{E}[|\mathbf{r}_k|^2]$ wherever necessary, and setting $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)| n^{1/2}K/\mu)/\mu nK\}$.

5

As we show in Appendix C, the update rule (5) resembles the linearized update rule of full batch GDA with added noise. In fact, for $n = 1$, $\mathbf{r}_k = 0$ and thus, we recover the rates of full-batch GDA. Expressing GDA-RR/SO (and later AS) as noisy full-batch GDA in this fashion is a central component of our unified analysis, and relies on the fact that $\sum_{i=1}^{n} \omega_{\tau_k(i)}(\mathbf{z}^*) = 0 \; \forall \; \tau_k \in \mathbb{S}_n$, which is specific to sampling without replacement. Comparing to SGDA with replacement, we note that sampling the components i.i.d. uniformly as $u(i) \sim \text{Uniform}([n])$ gives rise to an *additional* noise term $\alpha \mathbf{p}_k$ in the update rule, where $\mathbf{p}_k = \sum_{i=1}^{n} \omega_{u(i)}(\mathbf{z}^*)$ vanishes only in expecation, and has a variance of $\mathbb{E}[|\mathbf{p}_k|^2] = n\sigma_*^2$. Subsequently, the dominant noise term for SGDA updates is $O(\alpha^2 n \sigma_*^2)$ whereas that of GDA-RR/SO is $O(\alpha^4 n^3 \sigma_*^2)$, which qualitatively demonstrates the *implicit variance reduction* of sampling without replacement. As we shall see in the complete proof, this allows RR/SO to converge faster (for large enough K) by carefully selecting α. $\qquad\square$

Comparison with lower bounds: Since smooth strongly convex minimization is a special case of (4), the $\Omega(1/nK^2)$ lower bound established in prior works [48, 43] for smooth strongly convex minimization using GD with RR/SO also applies to GDA-RR/SO. Comparing with this lower bound, we note that the convergence rate of GDA-RR/SO is *nearly tight*, i.e., it differs from the lower bound only by an exponentially decaying term. In fact, for $K \geq 10\kappa^2 \log(n^{1/2}K)$, the convergence rate becomes $\tilde{O}(1/nK^2)$, which precisely matches the lower bound, modulo logarithmic factors.

Comparison with uniform sampling: Similarly, the $\Omega(1/nK)$ lower bound of SGD with replacement for smooth and strongly convex functions [44] also applies to SGDA with replacement. On the contrary, both GDA-RR and GDA-SO converge with a faster rate of $\tilde{O}(1/nK^2)$ when $K \geq 10\kappa^2 \log(n^{1/2}K)$. Thus, GDA-RR/SO provably outperform SGDA with replacement (modulo logarithmic factors) when $K \geq 10\kappa^2 \log(n^{1/2}K)$. As we show in Appendix C.2, the κ^2 dependence of this inequality cannot be improved for constant step-sizes. A similar argument also applies to stochastic PPM. To the best of our knowledge, the fastest known convergence rate for stochastic PPM is $O(1/nK)$ for minimizing smooth strongly convex functions [41]. Hence, Theorem 1 suggests that PPM-RR/SO enjoy a faster $\tilde{O}(1/nK^2)$ convergence rate *for both minimization and minimax optimization* when $K \geq 10\kappa^2 \log(n^{1/2}K)$.

3.3 Analysis in the Adversarial Shuffling Regime

We now consider without-replacement minimax optimization algorithms in an adversarial setting. We focus on a novel class of training-time attacks known as *data ordering attacks* proposed by Shumailov et al. [51]. These attacks differ from standard data-perturbation attacks [19] and exploit the fact that most implementations of stochastic gradient optimizers do not verify whether the permutation τ_k is truly sampled at random. Shumailov et al. [51] propose three distinct attack strategies, namely, *batch reordering*, which changes the order in which mini-batches are supplied to the algorithm, *batch reshuffling*, which changes the order in which individual data points are supplied, and *replacing* which prevents certain data points from being observed by the algorithm by consistently replacing them with other data points in the training set.

We analyze the convergence of without-replacement GDA and PPM under batch reshuffling attacks. To this end, we consider an adversarial modification of RR/SO where the permutations τ_k, instead of being sampled by the algorithm, are now selected by an adversary using a strategy unknown to the algorithm. We also the assume that, while choosing τ_k, the adversary is computationally unrestricted and has complete knowledge of all the components ω_i, the minimax point $\mathbf{z}^*$, and the iterates $\mathbf{z}_i^k$ observed so far. We call this setup *Adversarial Shuffling* (AS) and obtain convergence rates of GDA and PPM (named GDA-AS and PPM-AS) when solving (4). Thus, our analysis naturally holds for minimization, minimax optimization as well as finite-sum multiplayer games. The details are stated in Algorithms 1 and 2, respectively. Our last iterate convergence guarantees are deterministic and hold uniformly over any sequence of permutations $\tau_1, \ldots, \tau_K$ that the adversary can choose.

Theorem 2 (Convergence of GDA-AS and PPM-AS). *Consider Problem* (4) *for the μ-strongly monotone operator $\nu(\mathbf{z}) = 1/n \sum_{i=1}^{n} \omega_i(\mathbf{z})$ where each ω_i is l-Lipschitz, but not necessarily monotone. Let $\mathbf{z}^*$ denote the unique root of ν. Then, there exists a step-size $\alpha \leq \mu/5nl^2$ for which both GDA-AS and PPM-AS satisfy the following for any $K \geq 1$:*

$$\max_{\tau_1, \ldots, \tau_K \in \mathbb{S}_n} |\mathbf{z}_0^{K+1} - \mathbf{z}^*|^2 \leq 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 24\kappa^2\sigma_*^2 \log^3(|\nu(\mathbf{z}_0)|K/\mu)}{\mu^2 K^2} = \tilde{O}(e^{-K/5\kappa^2} + 1/K^2),$$

6

where κ, σ_*^2 are as defined in Theorem 1 and $\tau_1, \ldots, \tau_K$ are the permutations chosen by the adversary.

Convergence rates of IG and comparison with lower bounds: We note that the Incremental Gradient and the Incremental Proximal Point Methods, which do not shuffle the data, are a special case of GDA-AS/PPM-AS with $\tau_1, \ldots, \tau_K = id$. Thus, Theorem 2 also gives us convergence rates for GDA-IG/PPM-IG. Moreover, since GDA-AS generalizes GDA-IG and (4) covers minimization, the $\Omega(1/K^2)$ lower bound of IG established in prior works [48] for smooth strongly convex minimization also applies to GDA-AS. Thus, our obtained rate for GDA-AS is nearly tight and matches the lower bound (modulo logarithmic factors) for $K \geq 10\kappa^2 \log(K)$.

Effectiveness of batch reshuffling: When $K \geq 10\kappa^2 \log(K)$, $\tilde{O}(1/K^2)$ becomes the dominant term in the convergence rate of AS. This is worse than that of RR/SO by a factor of $1/n$ and causes a significant slowdown in convergence, since, in many applications, the dataset size n is much larger than K. Thus, our analysis justifies the effectiveness of batch reshuffling attacks in reducing model accuracy and increasing training time, which is empirically verified by Shumailov et al. [51].

4 RR for Two-Sided PŁ Objectives

We now analyze RR for a class of smooth nonconvex-nonconcave problems where the objective F satisfies a *two-sided Polyak Łojasiewicz inequality*, first proposed in Yang et al. [54]. We denote this function class as 2PŁ and formally state the assumption as follows.

Assumption 3 (Two-sided Polyak Łojasiewicz Inequality or 2PŁ condition)**.** *For any $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}, \mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}$, the sets $\arg\max_{\tilde{\mathbf{y}}} F(\mathbf{x}, \tilde{\mathbf{y}})$ and $\arg\min_{\tilde{\mathbf{x}}} F(\tilde{\mathbf{x}}, \mathbf{y})$ are non-empty. Furthermore, there exist positive constants μ_1, μ_2 such that F satisfies the following:*

$$|\nabla_\mathbf{x} F(\mathbf{x}, \mathbf{y})|^2 \geq 2\mu_1 [F(\mathbf{x}, \mathbf{y}) - \min_{\tilde{\mathbf{x}} \in \mathbb{R}^{d_\mathbf{x}}} F(\tilde{\mathbf{x}}, \mathbf{y})], \quad |\nabla_\mathbf{y} F(\mathbf{x}, \mathbf{y})|^2 \geq 2\mu_2 [\max_{\tilde{\mathbf{y}} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \tilde{\mathbf{y}}) - F(\mathbf{x}, \mathbf{y})].$$

The 2PŁ condition is satisfied in several practical settings, including, but not limited to, robust least squares [12], imitation learning for linear quadratic regulators [14, 8], and various other problems in reinforcement learning and robust control [11, 8]. Clearly, any SC-SC function is 2PŁ. However, 2PŁ functions need not be SC-SC, or even convex-concave. We refer the readers to Yang et al. [54] for a detailed discussion of the 2PŁ class and its applications.

Analysis of RR for 2PŁ objectives is challenging not only due to nonconvexity-nonconcavity, but also because F may not have a unique minimax point. Indeed, as we demonstrate in Appendix E.1, it is possible to construct 2PŁ functions where the set of minimax points is an unbounded proper subset of $\mathbb{R}^d$. Hence, the notion of *gradient variance at the minimax point*, which we used in our earlier analyses, is no longer meaningful. To overcome this, we impose the following assumption.

Assumption 4 (Bounded Gradient Variance)**.** *There exists a positive constant σ such that the component gradient operators ω_i satisfy the following for any $\mathbf{z} \in \mathbb{R}^d$:*

$$1/n \sum_{i=1}^n |\omega_i(\mathbf{z}) - \nu(\mathbf{z})|^2 \leq \sigma^2.$$

4.1 Analysis of AGDA-RR and AGDA-AS

In order to establish the provable benefits of RR for smooth finite-sum minimax optimization of 2PŁ objectives, we propose the *Alternating Gradient Descent Ascent with Random Reshuffling* (AGDA-RR) algorithm. AGDA-RR achieves near-optimal convergence guarantees for 2PŁ objectives by combining RR with *alternating updates* [17, 57, 3] and *timescale separation* [29, 15, 16], two ideas that have been very useful for improving convergence and stability in nonconvex-nonconcave minimax optimization. Within each epoch $k \in [K]$, AGDA-RR uniformly samples a random permutation τ_k, makes one full pass over the dataset, and performs gradient descent (with RR) updates for the variable $\mathbf{x}$ using the permutation τ_k. This is followed by sampling another permutation π_k and performing gradient ascent (with RR) updates for $\mathbf{y}$ using the permutation π_k. The detailed procedure is stated in Algorithm 3. We also analyze a variant of AGDA-RR in the adversarial shuffling setting, which we call AGDA-AS. The procedure, as described in Algorithm 4, is almost identical to AGDA-RR, except that the permutations τ_k and π_k are chosen by an adversary.

7

Algorithm 3: AGDA-RR	**Algorithm 4:** AGDA-AS
Input : Number of epochs K, step-sizes $\alpha, \beta > 0$, and initialization $(\mathbf{x}_0, \mathbf{y}_0)$	**Input** : Number of epochs K, step-sizes $\alpha, \beta > 0$, and initialization $(\mathbf{x}_0, \mathbf{y}_0)$
Initialize $(\mathbf{x}_0^1, \mathbf{y}_0^1) \leftarrow (\mathbf{x}_0, \mathbf{y}_0)$	Initialize $(\mathbf{x}_0^1, \mathbf{y}_0^1) \leftarrow (\mathbf{x}_0, \mathbf{y}_0)$
for $k \in [K]$ **do**	**for** $k \in [K]$ **do**
  Sample a permutation $\tau_k \in \mathbb{S}_n$	 Adversary chooses a permutation $\tau_k \in \mathbb{S}_n$
  for $i \in [n]$ **do**	 **for** $i \in [n]$ **do**
    $\mathbf{x}_i^k \leftarrow \mathbf{x}_{i-1}^k - \alpha \nabla_\mathbf{x} f_{\tau_k(i)}(\mathbf{x}_{i-1}^k, \mathbf{y}_0^k)$	 $\mathbf{x}_i^k \leftarrow \mathbf{x}_{i-1}^k - \alpha \nabla_\mathbf{x} f_{\tau_k(i)}(\mathbf{x}_{i-1}^k, \mathbf{y}_0^k)$
  end	 **end**
  Sample a permutation $\pi_k \in \mathbb{S}_n$	 Adversary chooses a permutation $\pi_k \in \mathbb{S}_n$
  for $i \in [n]$ **do**	 **for** $i \in [n]$ **do**
    $\mathbf{y}_i^k \leftarrow \mathbf{y}_{i-1}^k + \beta \nabla_\mathbf{y} f_{\pi_k(i)}(\mathbf{x}_n^k, \mathbf{y}_{i-1}^k)$	 $\mathbf{y}_i^k \leftarrow \mathbf{y}_{i-1}^k + \beta \nabla_\mathbf{y} f_{\pi_k(i)}(\mathbf{x}_n^k, \mathbf{y}_{i-1}^k)$
  end	 **end**
  $(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1}) \leftarrow (\mathbf{x}_n^k, \mathbf{y}_n^k)$	 $(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1}) \leftarrow (\mathbf{x}_n^k, \mathbf{y}_n^k)$
end	**end**

Before presenting a convergence analysis, we highlight that the absence of a unique minimax point prevents us from using the squared distance to the optimum as a Lyapunov function. To this end, we use the Lyapunov function $V_\lambda : \mathbb{R}^d \to \mathbb{R}$ which was previously suggested by Yang et al. [54]. We begin by first defining the *best response function* $\Phi : \mathbb{R}^d \to \mathbb{R}$ and its minimum Φ^* as follows,

$$\Phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y}), \quad \Phi^* = \min_{\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}} \Phi(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}} \max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y}).$$

Assumption 3 ensures that Φ is well defined and the existence of a global minimax point guarantees that Φ^* is finite. Subsequently, for any $\lambda > 0$, we define the Lyapunov function V_λ as

$$V_\lambda(\mathbf{x}, \mathbf{y}) = [\Phi(\mathbf{x}) - \Phi^*] + \lambda[\Phi(\mathbf{x}) - F(\mathbf{x}, \mathbf{y})].$$

By definition of Φ, V_λ is non-negative for any $\lambda > 0$ and $V_\lambda(\mathbf{z}) = 0$ if and only if $\mathbf{z}$ is a minimax point of F. Hence, we present our convergence proofs for AGDA-RR and AGDA-AS in terms of V_λ.

Theorem 3 (Convergence of AGDA-RR/AS). *Let Assumptions 1, 3, and 4 hold and let $\eta = {}^{73l^2}/_{2\mu_2^2}$. Then, there exists a step-size $\alpha \leq {}^1/_{5\eta nl}$ such that for $\beta = \eta\alpha$, AGDA-RR satisfies the following for $\lambda = {}^1/_{10}$ and any $K \geq 1$:*

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq e^{-K/365\kappa^3} V_\lambda(\mathbf{z}_0) + \frac{\mu_1 + c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0) n^{1/2} K)}{\mu_1 n K^2} = \tilde{O}(e^{-K/365\kappa^3} + {}^1/_{nK^2}),$$

where $\kappa = \max\{{}^l/_{\mu_1}, {}^l/_{\mu_2}\}$ and $c > 0$ is a constant independent of $\kappa, \mu_1, \mu_2, \sigma^2$. Under the same setting, AGDA-AS satisfies the following ($\hat{c} > 0$ is a constant independent of $\kappa, \mu_1, \mu_2, \sigma^2$):

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \leq e^{-K/365\kappa^3} V_\lambda(\mathbf{z}_0) + \frac{\mu_1 + \hat{c}\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0) K)}{\mu_1 K^2} = \tilde{O}(e^{-K/365\kappa^3} + {}^1/_{K^2}),$$

where $\tau_1, \pi_1 \ldots, \tau_K, \pi_K$ are the permutations chosen by the adversary.

Convergence to a Saddle Point: As demonstrated in Appendix E.4, the convergence guarantee of Theorem 3, which is presented in terms of V_λ, can be easily translated into an equivalent convergence guarantee in terms of $\text{dist}(\mathbf{z}, \mathcal{Z}^*)^2$, where $\mathcal{Z}^*$ denotes the set of saddle points of F. In particular, Theorem 3 implies the following convergence guarantee for AGDA-RR:

$$\mathbb{E}[\text{dist}(\mathbf{z}_0^{K+1}, \mathcal{Z}^*)^2] = \tilde{O}(e^{-K/365\kappa^3} + {}^1/_{nK^2}),$$

as well as the following convergence rate for AGDA-AS:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} \text{dist}(\mathbf{z}_0^{K+1}, \mathcal{Z}^*)^2 = \tilde{O}(e^{-K/365\kappa^3} + {}^1/_{K^2}),$$

Comparison with lower bounds: Strongly convex minimization is a special case of 2PŁ minimax optimization, since minimizing the strongly convex function f is equivalent to minimax optimization of the 2PŁ function $F(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \langle \mathbf{y}, \mathbf{y} \rangle$. In fact, the $\mathbf{x}$ iterates of AGDA-RR for F are exactly

8

| (a) GDA | (b) PPM | (c) AGDA |
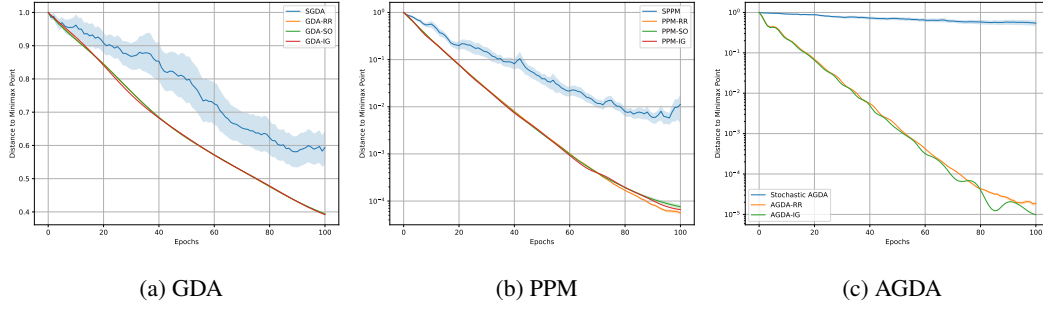
Figure 2: Relative distance of the epoch iterates from the global minimax point (i.e. $|\mathbf{z}_0^k - \mathbf{z}^*|^2/|\mathbf{z}_0 - \mathbf{z}^*|^2$ vs k). The solid lines are the average over 50 runs and the shaded regions are 95% confidence intervals. The y-axis of 3a is on a linear scale whereas that of 3b and 3c is on a logarithmic scale.

that of GD with RR for f. Hence, the $\Omega(1/nK^2)$ lower bound for strongly convex minimization using GD with RR also applies to AGDA-RR. Thus our convergence rate for AGDA-RR is nearly tight and matches the lower bound (modulo logarithmic factors) for $K \geq 730\kappa^3 \log(n^{1/2}K)$. Similarly, the Incremental Gradient version of AGDA is a special case of AGDA-AS with $\tau_1, \pi_1, \ldots, \tau_K, \pi_K = id$ and hence, AGDA-AS is nearly tight and matches the $\Omega(1/K^2)$ lower bound (modulo logarithmic factors) for $K \geq 730\kappa^3 \log(K)$.

Comparison with stochastic AGDA: Similarly, the $\Omega(1/nK)$ lower bound of SGD with replacement also holds for the Stochastic AGDA algorithm [54], which samples the component functions with replacement and performs two-timescale alternating updates similar to AGDA-RR. Hence, Theorem 3 demonstrates that AGDA-RR provably outperforms stochastic AGDA when $K \geq 730\kappa^3 \log(n^{1/2}K)$.

Bounded iterate assumption Assumption 4 is also used in analyzing RR for PŁ function minimization [34]. In this setting, an alternative *bounded iterate assumption*, which assumes that all the iterates $\mathbf{z}_i^k$ lie within a compact set, has also been used [1]. As shown in Appendix E.2, our proof of Theorem 3 easily adapts to this assumption. In the absence of either assumption, Li et al. [28] use time-varying step-sizes to obtain *asymptotic* $O(1/K^2)$ rates for RR on PŁ (and more generally for KŁ) minimization.

5 Experiments

We evaluate our theoretical results by benchmarking on finite-sum SC-SC quadratic minimax games. This class of problems appears in several applications such as reinforcement learning [11], robust regression, [12] and online learning [26]. The objective F and the components f_i are given by:

$$F(\mathbf{x}, \mathbf{y}) = 1/n \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{x}^T \mathbf{B}\mathbf{y} - \frac{1}{2}\mathbf{y}^T \mathbf{C}\mathbf{y},$$

$$f_i(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}_i\mathbf{x} + \mathbf{x}^T \mathbf{B}_i\mathbf{y} - \frac{1}{2}\mathbf{y}^T \mathbf{C}_i\mathbf{y} - \mathbf{u}_i^T \mathbf{x} - \mathbf{v}_i^T \mathbf{y},$$

where $\mathbf{A}$ and $\mathbf{C}$ are strictly positive definite. We generate the components f_i randomly, such that $\sum_{i=1}^{n} \mathbf{u}_i = \sum_{i=1}^{n} \mathbf{v}_i = 0$ and the expected singular values of $\mathbf{B}$ are larger than that of $\mathbf{A}$ and $\mathbf{C}$. This ensures that the bilinear coupling term $\mathbf{x}^T \mathbf{B}\mathbf{y}$ is sufficiently strong, since a weak coupling practically reduces to quadratic minimization, which has already been investigated in prior works. Finally, to investigate how the presence of nonconvex-nonconcave components impacts convergence, a few randomly chosen f_i's are allowed to be nonconvex-nonconcave quadratics. For each algorithm analyzed in the text, we benchmark sampling without replacement against uniform sampling by running each method for 100 epochs using constant step-sizes that are selected independently for each method via grid search. Further details regarding the setup is discussed in Appendix G.

We present our results in Figure 2, where we plot the relative distance of the epoch iterates from the minimax point, defined as $|\mathbf{z}_0^k - \mathbf{z}^*|^2/|\mathbf{z}_0 - \mathbf{z}^*|^2$, averaged over 50 independent runs. In agreement with our theoretical findings, sampling without replacement consistently outperforms uniform sampling

9

(a) GDA	(b) PPM	(c) AGDA
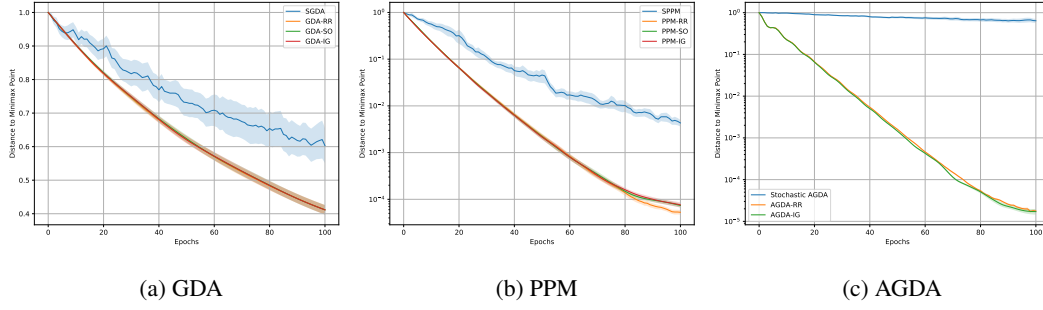
Figure 3: Convergence of GDA, PPM and AGDA averaged over 20 random instances. Shaded regions represent 95% confidence intervals.

across all three setups. Furthermore, to demonstrate that our observations are not particular to one specific instance, we repeat the experiment for 20 independently sampled quadratic games, and for each instance, perform 5 independent runs of each algorithm and plot the average relative distance of the epoch iterates from the minimax point. The results, presented in Figure 3, substantiates the superior convergence of sampling without replacement across multiple instances.

6   Conclusion

We derived near optimal convergence rates for several without-replacement stochastic gradient algorithms for finite-sum minimax optimization, and demonstrated that they converge faster than algorithms that use uniform sampling. We considered two problem classes, strongly convex-strongly concave problems (generalized to unconstrained strongly monotone variational inequalities) and nonconvex-nonconcave problems with two-sided PŁ objectives. We also formally defined *adversarial shuffling*, where an attacker can control the order in which data points are supplied to the optimizer, and analyzed minimax optimization in this regime. Interesting future directions include the analysis of inexact proximal point methods, more general function classes, and time-varying step-sizes.

Acknowledgements and Disclosure of Funding

Michael Muehlebach and Bernhard Schölkopf thank the German Research Foundation and the Branco Weiss Fellowship, administered by ETH Zurich, for the generous support.

References

[1] K. Ahn, C. Yun, and S. Sra. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. In *Advances in Neural Information Processing Systems*, 2020.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

[3] J. P. Bailey, G. Gidel, and G. Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Proceedings of the Conference on Learning Theory*, 2020.

[4] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade: Second Edition*, 2012.

[5] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Report LIDS - 2848, Laboratory for Information and Decision Systems, MIT*, 2011.

[6] J. Bose, G. Gidel, H. Berard, A. Cianflone, P. Vincent, S. Lacoste-Julien, and W. Hamilton. Adversarial example games. In *Advances in Neural Information Processing Systems*, 2020.

10

[7] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the Symposium on Learning and Data Science*, 2009.

[8] Q. Cai, M. Hong, Y. Chen, and Z. Wang. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv:1901.03674*, 2019.

[9] W. G. Cochran. *Sampling Techniques*. Wiley, 1977.

[10] C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the ACM SIGACT Symposium on the Theory of Computing*, 2021.

[11] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *Proceedings of the International Conference on Machine Learning*, 2017.

[12] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 1997.

[13] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[14] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the International Conference on Machine Learning*, 2018.

[15] T. Fiez and L. J. Ratliff. Local convergence analysis of gradient descent ascent with finite timescale separation. In *International Conference on Learning Representations*, 2021.

[16] T. Fiez, L. Ratliff, E. Mazumdar, E. Faulkner, and A. Narang. Global convergence to local min-max equilibrium in classes of nonconvex zero-sum games. In *Advances in Neural Information Processing Systems*, 2021.

[17] G. Gidel, R. A. Hemmat, M. Pezeshki, R. L. Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[20] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *Proceedings of the International Conference on Machine Learning*, 2019.

[21] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 2019.

[22] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Convergence rate of incremental gradient and incremental Newton methods. *SIAM Journal on Optimization*, 2019.

[23] J. Haochen and S. Sra. Random shuffling beats SGD after finite epochs. In *Proceedings of the International Conference on Machine Learning*, 2019.

[24] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

[25] P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Proceedings of the Conference on Learning Theory*, 2019.

[26] W. M. Koolen, A. Malek, and P. L. Bartlett. Efficient minimax strategies for square loss games. In *Advances in Neural Information Processing Systems*, 2014.

11

[27] Z. Lai and L.-H. Lim. Recht-Re noncommutative arithmetic-geometric mean conjecture is false. In *Proceedings of the International Conference on Machine Learning*, 2020.

[28] X. Li, A. Milzarek, and J. Qiu. Convergence of random reshuffling under the Kurdyka-Łojasiewicz inequality. *arXiv:2110.04926*, 2021.

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

[30] M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 1994.

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

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2018.

[33] C. Maheshwari, C.-Y. Chiu, E. Mazumdar, S. Sastry, and L. Ratliff. Zeroth-order methods for convex-concave minmax problems: applications to decision-dependent risk minimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2022.

[34] K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, 2020.

[35] K. Mishchenko, A. Khaled, and P. Richtárik. Proximal and federated random reshuffling. In *Proceedings of the International Conference on Machine Learning*, 2022.

[36] D. Nagaraj, P. Jain, and P. Netrapalli. SGD without replacement: sharper rates for general smooth convex functions. In *Proceedings of the International Conference on Machine Learning*, 2019.

[37] A. Nedic and D. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. In *Proceedings of the IEEE Conference on Decision and Control*, 1999.

[38] J. v. Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

[39] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. v. Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 2021.

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

[41] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 2018.

[42] A. Rajeswaran, I. Mordatch, and V. Kumar. A game theoretic framework for model based reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020.

[43] S. Rajput, A. Gupta, and D. Papailiopoulos. Closing the convergence gap of SGD without replacement. In *Proceedings of the International Conference on Machine Learning*, 2020.

[44] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the International Conference on Machine Learning*, 2012.

[45] B. Recht and C. Re. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In *Proceedings of the Conference on Learning Theory*, 2012.

12

[46] J. Rice. *Mathematical Statistics and Data Analysis*. Wadsworth, 1988.

[47] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 1976.

[48] I. Safran and O. Shamir. How good is SGD with random shuffling? In *Proceedings of the Conference on Learning Theory*, 2020.

[49] I. Safran and O. Shamir. Random shuffling beats SGD only after many epochs on ill-conditioned problems. In *Advances in Neural Information Processing Systems*, 2021.

[50] G. Scutari, D. P. Palomar, F. Facchinei, and J.-S. Pang. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 2010.

[51] I. Shumailov, Z. Shumaylov, D. Kazhdan, Y. Zhao, N. Papernot, M. A. Erdogdu, and R. J. Anderson. Manipulating SGD with data ordering attacks. In *Advances in Neural Information Processing Systems*, 2021.

[52] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *Proceedings of the International Conference on Learning Representations*, 2018.

[53] T. H. Tran, L. M. Nguyen, and Q. Tran-Dinh. SMG: A shuffling gradient-based method with momentum. In *Proceedings of the International Conference on Machine Learning*, 2021.

[54] J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, 2020.

[55] Y. Yu, T. Lin, E. Mazumdar, and M. I. Jordan. Fast distributionally robust learning with variance reduced min-max optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2022.

[56] C. Yun, S. Rajput, and S. Sra. Minibatch vs local SGD with shuffling: Tight convergence bounds and beyond. In *International Conference on Learning Representations*, 2022.

[57] G. Zhang, P. Poupart, and Y. Yu. Optimality and stability in non-convex smooth games. *Journal of Machine Learning Research*, 2022.

[58] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, 2019.

13

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] Limitations such as structural assumptions on the function class and noise variance are clearly stated wherever appropriate.

   (c) Did you discuss any potential negative societal impacts of your work? [No] The work is purely of a theoretical nature and analyzes existing algorithms that are already widely prevalent.

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions are stated in Sections 2, 3 and 4.

   (b) Did you include complete proofs of all theoretical results? [Yes] A proof sketch of the main result is presented in Section 3.2 while full proofs are deferred to the appendix.

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] They are included in the supplementary material

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] .Key details are specified in Section 5 and further details are specified in the appendix.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All results are reported with 95% confidence intervals.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] .These details are provided in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

14

Appendix

Contents

1   **Introduction**   **1**

     1.1   Contributions   2

2   **Notation and Preliminaries**   **3**

3   **Analysis for Strongly Convex-Strongly Concave Objectives**   **4**

     3.1   Setting   4

     3.2   Analysis of RR/SO   5

     3.3   Analysis in the Adversarial Shuffling Regime   6

4   **RR for Two-Sided PŁ Objectives**   **7**

     4.1   Analysis of AGDA-RR and AGDA-AS   7

5   **Experiments**   **9**

6   **Conclusion**   **10**

A   **Notation**   **16**

B   **Useful Lemmas**   **16**

C   **Analysis of GDA without Replacement**   **18**

     C.1   Analysis of Full-batch GDA by Linearization   19

     C.2   A Unified Analysis of GDA-RR and GDA-SO   20

     C.3   Analysis of GDA-AS   28

D   **Analysis of PPM without Replacement**   **30**

     D.1   Analysis of Full-batch PPM by Linearization   31

     D.2   A Unified Analysis of PPM-RR and PPM-SO   32

     D.3   Analysis of PPM-AS   37

E   **Analysis of AGDA-RR and AGDA-AS**   **39**

     E.1   Properties of Two-sided PŁ Functions   39

     E.2   Analysis of AGDA-RR   40

     E.3   Analysis of AGDA-AS   48

     E.4   Convergence of AGDA-RR and AGDA-AS to a Saddle Point   51

F   **Literature Review**   **53**

G   **Experimental Details**   **54**

15

A   Notation

In addition to the existing notation introduced in Section 2, we introduce the matrix product notation $\prod_{i=1}^{n} \mathbf{M}_i = \mathbf{M}_1 \mathbf{M}_2 \ldots \mathbf{M}_n$ where $\mathbf{M}_1, \mathbf{M}_2, \ldots \mathbf{M}_n$ are square matrices of the same size. Since matrix multiplication is not commutative, the ordering of the indices is important. To this end, we denote the reverse ordered matrix product as follows:

$$\widetilde{\prod}_{i=1}^{n} \mathbf{M}_i = \mathbf{M}_n \mathbf{M}_{n-1} \ldots \mathbf{M}_1.$$

We highlight that the above product is not necessarily equal to $\prod_{i=1}^{n} \mathbf{M}_i$ due to the noncommutativity of matrix products. Furthermore, we define the empty (reverse) matrix product (i.e. the matrix product or reverse matrix product to be indexed over empty sets) to be the identity matrix $\mathbf{I}$, where the size of $\mathbf{I}$ is clear from the context. Finally, we use $\sigma_{\max}(\mathbf{A})$ to denote the maximum singular value of the matrix $\mathbf{A}$ (which is equal to its operator norm $|\mathbf{A}|$) and $\sigma_{\min}(\mathbf{A})$ to denote its minimum singular value.

B   Useful Lemmas

In this section, we present the proofs of some properties that are important for our convergence analysis. We begin with a lemma on the differentiability properties of Lipschitz and strongly monotone operators, which is repeatedly used in all our proofs.

Lemma B.1 (Differentiability properties of Lipschitz operators). *Let $\nu : \mathbb{R}^d \to \mathbb{R}^d$ be an l-Lipschitz continuous operator. Then, ν is Lebesgue almost everywhere differentiable, i.e., the Jacobian $\nabla \nu(\mathbf{z})$ is well defined everywhere outside of a set of Lebesgue measure zero and satisfies $|\nabla \nu(\mathbf{z})| \leq l$, wherever it is defined. Additionally, if ν is μ-strongly monotone, then ν has a unique root, and $\mathbf{v}^T \nabla \nu(\mathbf{z}) \mathbf{v} \geq \mu |\mathbf{v}|^2$ for any $\mathbf{v} \in \mathbb{R}^d$ and any $\mathbf{z} \in \mathbb{R}^d$ such that $\nabla \nu(\mathbf{z})$ is well defined.*

Proof. The first statement of this lemma is a direct consequence of Rademacher's Theorem which states that any Lipschitz continuous function defined on $\mathbb{R}^d$ is differentiable everywhere outside of a set of zero Lebesgue measure. The second statement is proved as follows:

Consider any $\mathbf{z} \in \mathbb{R}^d$ such that $\nabla \nu(\mathbf{z})$ is well defined and let $\mathbf{v} \in \mathbb{R}^d$ be arbitrary. By the Lipschitz continuity of ν, it follows that for any $t \in \mathbb{R}$:

$$|\nu(\mathbf{z} + t\mathbf{v}) - \nu(\mathbf{z})| \leq l \, |t\mathbf{v}| \,.$$

Rearranging and taking limits results in,

$$\lim_{t \to 0} \left| \frac{\nu(\mathbf{z} + t\mathbf{v}) - \nu(\mathbf{z})}{t} \right| \leq l \, |\mathbf{v}| \,.$$

By the continuity of norms and the definition of the Jacobian, it follows that $|\nabla \nu(\mathbf{z})\mathbf{v}| \leq l \, |\mathbf{v}| \; \forall \; \mathbf{v} \in \mathbb{R}^d$, which further implies that $|\nabla \nu(\mathbf{z})| \leq l$. This concludes the proof of the second statement.

For the proof of the third statement, we consider the operator $\zeta(\mathbf{z}) = \mathbf{z} - \frac{\mu}{l^2} \nu(\mathbf{z})$ and observe that any fixed point of ζ is a root of ν and vice versa.

From the l-smoothness and μ-strong monotonicity of ν, we infer that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$:

$$|\zeta(\mathbf{z}_2) - \zeta(\mathbf{z}_1)|^2 = |\mathbf{z}_2 - \mathbf{z}_1|^2 - \frac{2\mu}{l^2} \langle \nu(\mathbf{z}_2) - \nu(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1 \rangle + \frac{\mu^2}{l^4} |\nu(\mathbf{z}_2) - \nu(\mathbf{z}_1)|^2$$

$$\leq (1 - \frac{\mu^2}{l^2}) |\mathbf{z}_2 - \mathbf{z}_1|^2 < |\mathbf{z}_2 - \mathbf{z}_1|^2 \,.$$

Thus, ζ is a contraction mapping, and has a unique fixed point $\mathbf{z}^*$ by the Banach Fixed Point Theorem, which implies that $\mathbf{z}^*$ is the unique root of ν. Finally, consider any $\mathbf{z} \in \mathbb{R}^d$ such that $\nabla \nu(\mathbf{z})$ is well defined and let $\mathbf{v} \in \mathbb{R}^d$ be arbitrary. By the strong monotonicity of ν, it follows that for any $t \in \mathbb{R}$:

$$\langle \nu(\mathbf{z} + t\mathbf{v}) - \nu(\mathbf{z}), t\mathbf{v} \rangle \geq \mu \, |t\mathbf{v}|^2 \,.$$

Rearranging and taking limits results in

$$\left\langle \mathbf{v}, \lim_{t \to 0} \frac{\nu(\mathbf{z} + t\mathbf{v}) - \nu(\mathbf{z})}{t} \right\rangle \geq \mu \, |\mathbf{v}|^2 \,.$$

16

By definition of the Jacobian and the continuity of inner products, it follows that $\mathbf{v}^T \nabla \nu(\mathbf{z}) \mathbf{v} \geq \mu |\mathbf{v}|^2$. $\qquad \square$

We now present a lemma on the properties of the gradient operator of a strongly convex-strongly concave function, which was earlier stated as Lemma 1 in Section 3.1. This lemma establishes the equivalence between the minimax optimization problem for a smooth and strongly convex-strongly concave function F and the root finding problem for its corresponding gradient operator ν.

Lemma B.2 (Equivalence of minimax optimization and root finding)**.** *Let $F : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}^{d_\mathbf{y}} \to \mathbb{R}$ be an l-smooth and μ-strongly convex-strongly concave function and let $\nu(\mathbf{x}, \mathbf{y}) = [\nabla_\mathbf{x} F(\mathbf{x}, \mathbf{y}), -\nabla_\mathbf{y} F(\mathbf{x}, \mathbf{y})]$. Then ν is l-Lipschitz continuous, μ-strongly monotone and has a unique root $\mathbf{z}^*$, which is also the unique global minimax point of F.*

Proof. The l-Lipschitz continuity of ν follows by definition of the l-smoothness of F, as described in Assumption 1. To establish μ-strong monotonicity, we proceed as follows:

Consider any $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1)$ and $\mathbf{z}_2 = (\mathbf{x}_2, \mathbf{y}_2)$ in $\mathbb{R}^d$. Since F is μ-strongly convex in $\mathbf{x}$ and μ-strongly concave in $\mathbf{y}$, it follows that:

$$F(\mathbf{x}_2, \mathbf{y}_1) - F(\mathbf{x}_1, \mathbf{y}_1) \geq \langle \nabla_\mathbf{x} F(\mathbf{x}_1, \mathbf{y}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle + \frac{\mu}{2} |\mathbf{x}_2 - \mathbf{x}_1|^2,$$

$$F(\mathbf{x}_1, \mathbf{y}_2) - F(\mathbf{x}_2, \mathbf{y}_2) \geq \langle \nabla_\mathbf{x} F(\mathbf{x}_2, \mathbf{y}_2), \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\mu}{2} |\mathbf{x}_2 - \mathbf{x}_1|^2,$$

$$F(\mathbf{x}_2, \mathbf{y}_2) - F(\mathbf{x}_2, \mathbf{y}_1) \geq \langle \nabla_\mathbf{y} F(\mathbf{x}_2, \mathbf{y}_2), \mathbf{y}_2 - \mathbf{y}_1 \rangle + \frac{\mu}{2} |\mathbf{y}_2 - \mathbf{y}_1|^2,$$

$$F(\mathbf{x}_1, \mathbf{y}_1) - F(\mathbf{x}_1, \mathbf{y}_2) \geq \langle \nabla_\mathbf{y} F(\mathbf{x}_1, \mathbf{y}_1), \mathbf{y}_1 - \mathbf{y}_2 \rangle + \frac{\mu}{2} |\mathbf{y}_2 - \mathbf{y}_1|^2.$$

Adding all four inequalities, and substituting the definition of ν, we get,

$$\langle \nu(\mathbf{z}_2) - \nu(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1 \rangle \geq \mu |\mathbf{z}_2 - \mathbf{z}_1|^2.$$

To prove the second part of this lemma, we define $\Phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y})$. Furthermore, we note that Φ is strongly convex and by Danskin's Theorem, $\nabla \Phi(\mathbf{x}) = \nabla_\mathbf{x} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ where $\mathbf{y}^*(\mathbf{x}) \in \mathbb{R}^{d_\mathbf{y}}$ is the unique solution of $F(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) = \max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y})$, with uniqueness being guaranteed by the strong concavity of F in $\mathbf{y}$. Since ν is l-Lipschitz continuous and μ-strongly monotone, it is guaranteed to have a unique root $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ by Lemma B.1, which implies that $\nabla_\mathbf{x} F(\mathbf{x}^*, \mathbf{y}^*) = \nabla_\mathbf{y} F(\mathbf{x}^*, \mathbf{y}^*) = 0$. Since F is strongly concave in $\mathbf{y}$, this further implies that $\mathbf{y}^*$ is the unique maximizer of $F(\mathbf{x}^*, \mathbf{y})$. Moreover, by Danskin's Theorem, $\nabla \Phi(\mathbf{x}^*) = \nabla_\mathbf{x} F(\mathbf{x}^*, \mathbf{y}^*) = 0$ which implies that $\mathbf{x}^*$ is the unique minimizer of Φ, since Φ is strongly convex. Thus, by definition of a global minimax point, we conclude that $\mathbf{z}^*$ is the unique global minimax point of F. $\qquad \square$

We conclude this section with the proof of a standard result in statistics on the variance of the population mean estimator under simple random sampling without replacement [9, 46]. This property constitutes a central component of our convergence proofs and allows us to sharply bound the noise in the iterates of RR and SO. This lemma has also been used in Mishchenko et al. [34] to analyze RR/SO for minimization.

Lemma B.3 (Properties of without-replacement sample means : Lemma 1 in Mishchenko et al. [34])**.** *Let τ be a uniformly sampled random permutation of $[n]$, and let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ be arbitrary vectors that are **independent of** τ, with sample mean $\mathbf{m} = 1/n \sum_{i=1}^{n} \mathbf{v}_i$ and sample variance $\sigma^2 = 1/n \sum_{i=1}^{n} |\mathbf{v}_i - \mathbf{m}|^2$. For any $i \in [n]$, let the without-replacement sample average be defined as $\hat{\mathbf{m}}_\tau^{(i)} = 1/i \sum_{j=1}^{i} \mathbf{v}_{\tau(j)}$. Then, $\mathbb{E}[\hat{\mathbf{m}}_\tau^{(i)}] = \mathbf{m}$ and $\mathbb{E}[|\hat{\mathbf{m}}_\tau^{(i)} - \mathbf{m}|^2] = \frac{n-i}{n-1} \frac{\sigma^2}{i}$, where the expectation is taken over the uniform random permutation τ.*

Proof. Since $\mathbf{v}_1, \dots, \mathbf{v}_n$ is independent of the uniformly sampled permutation τ, it follows that for any $j \in [n]$,

$$\mathbb{E}[\mathbf{v}_{\tau(j)}] = \sum_{k=1}^{n} \mathbb{P}_{\text{Uniform}}[\tau(j) = k] \mathbf{v}_k = \sum_{k=1}^{n} \frac{(n-1)!}{n!} \mathbf{v}_k = 1/n \sum_{k=1}^{n} \mathbf{v}_k = \mathbf{m}.$$

17

Hence, by linearity of expectations,

$$\mathbb{E}[\hat{\mathbf{m}}_\tau^{(i)}] = 1/i \sum_{j=1}^{i} \mathbb{E}[\mathbf{v}_{\tau(j)}] = \mathbf{m}.$$

Similarly, we can show that for any $j \in [n]$,

$$\mathbb{E}[|\mathbf{v}_{\tau(j)} - \mathbf{m}|^2] = \sum_{k=1}^{n} \frac{(n-1)!}{n!} |\mathbf{v}_k - \mathbf{m}|^2 = 1/n \sum_{k=1}^{n} |\mathbf{v}_k - \mathbf{m}|^2 = \sigma^2.$$

We now consider any $j_1, j_2 \in [n]$ such that $j_1 \neq j_2$ and infer that,

$$\mathbb{E}[\langle \mathbf{v}_{\tau(j_1)} - \mathbf{m}, \mathbf{v}_{\tau(j_2)} - \mathbf{m}\rangle] = \sum_{1 \leq k_1 \neq k_2 \leq n} \mathbb{P}_{\text{Uniform}}\left[\tau(j_1) = k_1, \tau(j_2) = k_2\right] \langle \mathbf{v}_{k_1} - \mathbf{m}, \mathbf{v}_{k_2} - \mathbf{m}\rangle$$

$$= \frac{(n-2)!}{n!} \sum_{k_1=1}^{n} \sum_{1 \leq k_2 \neq k_1 \leq n} \langle \mathbf{v}_{k_1} - \mathbf{m}, \mathbf{v}_{k_2} - \mathbf{m}\rangle$$

$$= \frac{1}{n(n-1)} \sum_{k_1=1}^{n} \left[\left(\sum_{k_2=1}^{n} \langle \mathbf{v}_{k_1} - \mathbf{m}, \mathbf{v}_{k_2} - \mathbf{m}\rangle\right) - |\mathbf{v}_{k_1} - \mathbf{m}|^2\right]$$

$$= \frac{1}{n(n-1)} \sum_{k_1=1}^{n} \left\langle \mathbf{v}_{k_1} - \mathbf{m}, \sum_{k_2=1}^{n} \mathbf{v}_{k_2} - \mathbf{m}\right\rangle - \frac{1}{n(n-1)} \sum_{k_1=1}^{n} |\mathbf{v}_{k_1} - \mathbf{m}|^2$$

$$= -\frac{\sigma^2}{n-1}.$$

From the above identities, we conclude that,

$$\mathbb{E}[|\hat{\mathbf{m}}_\tau^{(i)} - \mathbf{m}|^2] = \mathbb{E}[|1/i \sum_{j=1}^{i} \mathbf{v}_{\tau(j)} - \mathbf{m}|^2]$$

$$= 1/i^2 \mathbb{E}[|\sum_{j=1}^{i} \mathbf{v}_{\tau(j)} - \mathbf{m}|^2]$$

$$= 1/i^2 \sum_{j=1}^{i} \sum_{k=1}^{i} \mathbb{E}[\langle \mathbf{v}_{\tau(j)} - \mathbf{m}, \mathbf{v}_{\tau(k)} - \mathbf{m}\rangle]$$

$$= 1/i^2 \left(\sum_{j=1}^{i} \mathbb{E}[|\mathbf{v}_{\tau(j)} - \mathbf{m}|^2] + \sum_{1 \leq j \neq k \leq i} \mathbb{E}[\langle \mathbf{v}_{\tau(j)} - \mathbf{m}, \mathbf{v}_{\tau(k)} - \mathbf{m}\rangle]\right)$$

$$= 1/i^2 [i\sigma^2 - \frac{i(i-1)}{n-1}\sigma^2]$$

$$= \frac{n-i}{n-1} \frac{\sigma^2}{i}.$$

$\square$

C Analysis of GDA without Replacement

In this section, we present the proofs of convergence for simultaneous Gradient Descent Ascent without replacement (i.e. GDA-RR/SO/AS) for solving the finite-sum strongly monotone root finding problem. Our proofs of the same rely on two key techniques: 1. exploiting the Lipschitz continuity of the components ω_i to *linearize* the dynamics of GDA without replacement about $\mathbf{z}^*$ (where $\mathbf{z}^*$ denotes the unique root of ν), 2. using Lemma B.3 to control the noise in the iterates of GDA without replacement.

For ease of exposition, we first illustrate the linearization technique by presenting a linearization-based proof of the well known exponential convergence rate of full-batch GDA for solving the root

18

finding problem under Lipschitz continuity and strong monotonicity assumptions. This serves as a prelude for our unified proof of convergence of GDA-RR and GDA-SO for solving the root finding problem for finite-sum strongly monotone operators. We conclude this section by presenting a proof of convergence of GDA-AS, which demonstrates how our techniques for analyzing RR and SO can be easily adapted to the adversarial shuffling setup.

C.1 Analysis of Full-batch GDA by Linearization

Theorem C.1 (Convergence of full-batch GDA). *Consider Problem* (4) *for the l-Lipschitz and μ-strongly monotone operator $\nu : \mathbb{R}^d \to \mathbb{R}^d$ and let $\mathbf{z}^*$ denote the unique root of ν. For any step-size $\alpha > 0$, the iterates $\mathbf{z}_k$ of full-batch GDA satisfy the following recurrence:*

$$\mathbf{z}_{k+1} - \mathbf{z}^* = (\mathbf{I} - \alpha \mathbf{M}_k)(\mathbf{z}_k - \mathbf{z}^*),$$

where the spectral norm of $\mathbf{I} - \alpha \mathbf{M}_k$ is bounded as $|\mathbf{I} - \alpha \mathbf{M}_k| \leq (1 - 2\alpha\mu + \alpha^2 l^2)^{1/2}$. Consequently, setting $\alpha = \mu/l^2$ gives us the following last iterate convergence guarantee:

$$|\mathbf{z}_{K+1} - \mathbf{z}^*|^2 \leq (1 - 1/\kappa^2)^K |\mathbf{z}_0 - \mathbf{z}^*|^2,$$

where $\kappa = l/\mu$. Furthermore, the exponential convergence rate of $1 - 1/\kappa^2$ is optimal for GDA up to constant factors.

Proof. The iterates of full-batch GDA with step-size α have the following update rule:

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \alpha\nu(\mathbf{z}_k).$$

Interpreting the time evolution of $\mathbf{z}_k$ as a discrete-time dynamical system, we analyze the convergence of GDA by applying *linearization*, which is a standard technique in the analysis of discrete and continuous-time dynamical systems. To this end, we proceed as follows.

Let $g : [0, 1] \to \mathbb{R}^d$ be defined as $g(t) = \nu(t\mathbf{z}_k + (1-t)\mathbf{z}^*)$. We note that the Lipschitz continuity of ν implies the Lipschitz continuity of g. As a result, the Fundamental Theorem of Calculus for Lebesgue Integrals applies to g (since Lipschitz continuous functions defined on a compact interval are also absolutely continuous). It follows that $g(1) = g(0) + \int_0^1 g'(t)\mathrm{d}t$ which results in the following expansion:

$$
\begin{aligned}
\nu(\mathbf{z}_k) &= \nu(\mathbf{z}^*) + \int_0^1 \nabla\nu(t\mathbf{z}_k + (1-t)\mathbf{z}^*)(\mathbf{z}_k - \mathbf{z}^*)\mathrm{d}t \\
&= \mathbf{M}_k(\mathbf{z}_k - \mathbf{z}^*),
\end{aligned}
$$

where $\nu(\mathbf{z}^*) = 0$ and $\mathbf{M}_k = \int_0^1 \nabla\nu(t\mathbf{z}_k + (1-t)\mathbf{z}^*)\mathrm{d}t$ is well defined because $\nabla\nu(\mathbf{z})$ is defined Lebesgue-almost everywhere by Lemma B.1.

Substituting the above expansion into the update rule of GDA gives us the following:

$$\mathbf{z}_{k+1} - \mathbf{z}^* = (\mathbf{I} - \alpha\mathbf{M}_k)(\mathbf{z}_k - \mathbf{z}^*). \tag{C.1}$$

We now proceed to obtain a convergence guarantee by upper bounding the spectral norm of $\mathbf{I} - \alpha\mathbf{M}_k$. Note that by Lemma B.1, $\nabla\nu(\mathbf{z})$ is well defined Lebesgue-almost everywhere, and wherever defined, satisfies $|\nabla\nu(\mathbf{z})| \leq l$ and $\mathbf{v}^T\nabla\nu(\mathbf{z})\mathbf{v} \geq \mu |\mathbf{v}|^2 \ \forall \ \mathbf{v} \in \mathbb{R}^d$. Using standard properties of Lebesgue integrals and inner products, we conclude that $|\mathbf{M}_k| \leq l$ and $\mathbf{v}^T\mathbf{M}_k\mathbf{v} \geq \mu |\mathbf{v}|^2 \ \forall \ \mathbf{v} \in \mathbb{R}^d$. Hence, for any $\mathbf{v} \in \mathbb{R}^d$:

$$
\begin{aligned}
|(\mathbf{I} - \alpha\mathbf{M}_k)\mathbf{v}|^2 &= |\mathbf{v}|^2 - \alpha\mathbf{v}^T\mathbf{M}_k\mathbf{v} - \alpha\mathbf{v}^T\mathbf{M}_k^T\mathbf{v} + |\mathbf{M}_k\mathbf{v}|^2 \\
&\leq (1 - 2\alpha\mu + \alpha^2 l^2) |\mathbf{v}|^2.
\end{aligned}
$$

Hence, $|\mathbf{I} - \alpha\mathbf{M}_k| \leq (1 - 2\alpha\mu + \alpha^2 l^2)^{1/2}$ for any $\alpha > 0$. It follows that,

$$|\mathbf{z}_{k+1} - \mathbf{z}^*|^2 \leq (1 - 2\alpha\mu + \alpha^2 l^2) |\mathbf{z}_k - \mathbf{z}^*|^2.$$

Setting $\alpha = \mu/l^2$, we obtain the following convergence guarantee:

$$|\mathbf{z}_{K+1} - \mathbf{z}^*|^2 \leq (1 - 1/\kappa^2)^K |\mathbf{z}_0 - \mathbf{z}^*|^2.$$

19

We now establish the optimality of the $1 - 1/\kappa^2$ exponential convergence rate by presenting a lower bound construction on a two-dimensional quadratic problem.

Consider any positive constants l, μ such that $l > \mu$ and the quadratic $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = \frac{\mu}{2}(x^2 - y^2) + (l - \mu)xy$. It is easy to see that f is l-smooth and μ-strongly convex strongly concave. Thus, by Lemma B.2, we infer that $\nu(\mathbf{z}) = \nu(x, y) = [\nabla_x f(x, y), -\nabla_y f(x, y)]$ is l-smooth and μ-strongly monotone. Furthermore, we also note that $\nu(\mathbf{z}) = \mathbf{M}\mathbf{z}$ where $\mathbf{M}$ is given by:

$$\mathbf{M} = \begin{bmatrix} \mu & l - \mu \\ \mu - l & \mu \end{bmatrix}.$$

Thus, it is easy to see that $\mathbf{z}^* = 0$ is the unique root of ν, where uniqueness is guaranteed by Lemma B.1. Moreover, for any step-size $\alpha > 0$, the iterates of GDA on f are given by $\mathbf{z}_{k+1} = (\mathbf{I} - \alpha \mathbf{M})\mathbf{z}_k$. Thus, by taking norms on both sides, we conclude that:

$$|\mathbf{z}_{k+1}|^2 = |(\mathbf{I} - \alpha\mathbf{M})\mathbf{z}_k|^2 = \left[1 - 2\alpha\mu + \alpha^2[\mu^2 + (l - \mu)^2]\right]|\mathbf{z}_k|^2$$

$$\geq (1 - \frac{\mu^2}{\mu^2 + (l - \mu)^2})|\mathbf{z}_k|^2 = (1 - \frac{1}{1 + (\kappa - 1)^2})|\mathbf{z}_k|^2,$$

where the inequality is obtained by minimizing the quadratic $1 - 2\alpha\mu + \alpha^2[\mu^2 + (l - \mu)^2]$ with respect to α.

We note that by Young's inequality, $\kappa^2 = [1 + (\kappa - 1)]^2 \leq 2[1 + (\kappa - 1)^2]$ and thus,

$$1 - \frac{1}{1 + (\kappa - 1)^2} \geq 1 - \frac{2}{\kappa^2},$$

which leads us to conclude that for any $\alpha > 0$, the iterates $\mathbf{z}_k$ of GDA on ν satisfy the following:

$$|\mathbf{z}_{k+1} - \mathbf{z}^*|^2 \geq (1 - 2/\kappa^2)|\mathbf{z}_k - \mathbf{z}^*|^2.$$

This establishes that the convergence rate of $1 - 1/\kappa^2$ is optimal for GDA up to constant factors. We contrast this with the smooth and strongly convex minimization setting where gradient descent enjoys a faster convergence rate of $1 - 1/\kappa$. $\qquad \square$

C.2 A Unified Analysis of GDA-RR and GDA-SO

Theorem C.2 (Convergence of GDA-RR/SO). *Consider Problem* (4) *for the μ-strongly monotone operator $\nu(\mathbf{z}) = 1/n \sum_{i=1}^n \omega_i(\mathbf{z})$ where each ω_i is l-Lipschitz, but not necessarily monotone. Let $\mathbf{z}^*$ denote the unique root of ν. Then, for any $\alpha \leq \frac{\mu}{5nl^2}$ and $K \geq 1$, the iterates of GDA-RR and GDA-SO satisfy the following:*

$$\mathbb{E}[|\mathbf{z}_0^{K+1} - \mathbf{z}^*|^2] \leq 2e^{-n\alpha\mu K}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{l^2\sigma_*^2\alpha^3 n^2 K}{\mu}.$$

Setting $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)|n^{1/2}K/\mu)/\mu nK\}$ results in the following expected last-iterate convergence guarantee for both GDA-RR and GDA-SO which holds for any $K \geq 1$:

$$\mathbb{E}[|\mathbf{z}_0^{K+1} - \mathbf{z}^*|^2] \leq 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 8\kappa^2\sigma_*^2\log^3(|\nu(\mathbf{z}_0)|n^{1/2}K/\mu)}{\mu^2 nK^2}$$

$$= \tilde{O}(e^{-K/5\kappa^2} + 1/nK^2).$$

Proof. Without loss of generality, we express the iterate-level update rule of GDA-RR and GDA-SO jointly as

$$\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k), \tag{C.2}$$

where $k \in [K]$ and $i \in [n]$. We note that for GDA-RR, $\tau_k \sim \text{Uniform}(\mathbb{S}_n)$ for every $k \in [K]$ (i.e. τ_k is a permutation of $[n]$ that is resampled uniformly in every epoch), whereas for GDA-SO, $\tau_k = \tau \; \forall k \in [K]$ where $\tau \sim \text{Uniform}(\mathbb{S}_n)$ (i.e. the permutation τ is uniformly sampled before the first epoch, then reused for all subsequent epochs).

20

As discussed in the proof sketch presented in Section 3.2, our proof begins by adapting the linearization technique illustrated in Theorem C.1 to GDA without replacement in order to derive a linearized epoch-level update rule of the form $\mathbf{z}_0^{k+1} - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^k - \mathbf{z}^*) + \alpha^2 \mathbf{r}_k$. We then use the Lipschitz continuity and strong monotonicity properties of ν to upper bound $|\mathbf{H}_k|$, and use Lemma B.3 to upper bound $\mathbb{E}[|\mathbf{r}_k|^2]$. Equipped with these bounds, the proof is concluded by unrolling the epoch-level update rule for K epochs and carefully choosing α as stated above.

Step 1: Linearized epoch-level update rule Our approach for this step is inspired by the following insight developed by prior works that have analyzed sampling without replacement for minimization: for small enough step-sizes, the epoch iterates $\mathbf{z}_0^k$ of gradient descent without replacement closely track the iterates of full-batch gradient descent. To this end, we treat GDA without replacement as a noisy version of full-batch GDA by focusing on the time evolution of the epoch iterates $\mathbf{z}_0^k$. At the same time, we wish to adapt the proof technique of Theorem C.1, which hinges on linearization about $\mathbf{z}^*$, to the analysis of GDA without replacement. This motivates us to perform the following decomposition of the stochastic gradients, where we linearize $\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k)$ about $\mathbf{z}_0^k$, and subsequently linearize $\omega_{\tau_k(i)}(\mathbf{z}_0^k)$ about $\mathbf{z}^*$:

$$\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k) = \omega_{\tau_k(i)}(\mathbf{z}^*) + [\omega_{\tau_k(i)}(\mathbf{z}_0^k) - \omega_{\tau_k(i)}(\mathbf{z}^*)] + [\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k) - \omega_{\tau_k(i)}(\mathbf{z}_0^k)]$$

$$= \omega_{\tau_k(i)}(\mathbf{z}^*) + \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_0^k + (1-t)\mathbf{z}^*)(\mathbf{z}_0^k - \mathbf{z}^*)\mathrm{d}t$$

$$+ \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_{i-1}^k + (1-t)\mathbf{z}_0^k)(\mathbf{z}_{i-1}^k - \mathbf{z}_0^k)\mathrm{d}t.$$

We define the matrices $\mathbf{M}_{\tau_k(i)}$ and $\mathbf{J}_{\tau_k(i)}$ as follows:

$$\mathbf{M}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_0^k + (1-t)\mathbf{z}^*)\mathrm{d}t,$$

$$\mathbf{J}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_{i-1}^k + (1-t)\mathbf{z}_0^k)\mathrm{d}t.$$

Repeating the same argument as in Theorem C.1, we use Lemma B.1 and the l-Lipschitz continuity of $\omega_i(\mathbf{z})$ $\forall i \in [n]$ to conclude that $\mathbf{M}_{\tau_k(i)}$ and $\mathbf{J}_{\tau_k(i)}$ are well defined and bounded with $\left|\mathbf{M}_{\tau_k(i)}\right| \leq l$ and $\left|\mathbf{J}_{\tau_k(i)}\right| \leq l$. It follows that

$$\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k) = \omega_{\tau_k(i)}(\mathbf{z}^*) + \mathbf{M}_{\tau_k(i)}(\mathbf{z}_0^k - \mathbf{z}^*) + \mathbf{J}_{\tau_k(i)}(\mathbf{z}_{i-1}^k - \mathbf{z}_0^k). \tag{C.3}$$

Substituting (C.3) into the update equation (C.2) for $\mathbf{z}_1^k$ gives us the following:

$$\mathbf{z}_1^k - \mathbf{z}^* = (\mathbf{I} - \alpha\mathbf{M}_{\tau_k(1)})(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha\mathbf{J}_{\sigma_k(1)}(\mathbf{z}_0^k - \mathbf{z}_0^k) - \alpha\omega_{\tau_k(1)}(\mathbf{z}^*)$$

$$= (\mathbf{I} - \alpha\mathbf{M}_{\tau_k(1)})(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha\omega_{\tau_k(1)}(\mathbf{z}^*).$$

Repeating the same for $\mathbf{z}_2^k$ yields,

$$\mathbf{z}_2^k - \mathbf{z}^* = \mathbf{z}_1^k - \mathbf{z}^* - \alpha\omega_{\tau_k(2)}(\mathbf{z}^*) - \alpha\mathbf{M}_{\tau_k(2)}(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha\mathbf{J}_{\tau_k(2)}(\mathbf{z}_1^k - \mathbf{z}_0^k)$$

$$= (\mathbf{I} - \alpha\mathbf{J}_{\tau_k(2)})(\mathbf{z}_1^k - \mathbf{z}^*) - \alpha(\mathbf{M}_{\tau_k(2)} - \mathbf{J}_{\tau_k(2)})(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha\omega_{\tau_k(2)}(\mathbf{z}^*)$$

$$= (\mathbf{I} - \alpha\mathbf{J}_{\tau_k(2)})[(\mathbf{I} - \alpha\mathbf{M}_{\tau_k(1)})(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha\omega_{\tau_k(1)}(\mathbf{z}^*)]$$

$$- \alpha(\mathbf{M}_{\tau_k(2)} - \mathbf{J}_{\tau_k(2)})(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha\omega_{\tau_k(2)}(\mathbf{z}^*)$$

$$= [(\mathbf{I} - \alpha\mathbf{J}_{\tau_k(2)})(\mathbf{I} - \alpha\mathbf{M}_{\tau_k(1)}) - \alpha(\mathbf{M}_{\tau_k(2)} - \mathbf{J}_{\tau_k(2)})](\mathbf{z}_0^k - \mathbf{z}^*)$$

$$- \alpha[\omega_{\tau_k(2)}(\mathbf{z}^*) + (\mathbf{I} - \alpha\mathbf{J}_{\tau_k(2)})\omega_{\tau_k(1)}(\mathbf{z}^*)]$$

$$= [\mathbf{I} - \alpha\mathbf{M}_{\tau_k(2)} - \alpha(\mathbf{I} - \alpha\mathbf{J}_{\tau_k(2)})\mathbf{M}_{\tau_k(1)}](\mathbf{z}_0^k - \mathbf{z}^*)$$

$$- \alpha[\omega_{\tau_k(2)}(\mathbf{z}^*) + (\mathbf{I} - \alpha\mathbf{J}_{\tau_k(2)})\omega_{\tau_k(1)}(\mathbf{z}^*)].$$

21

Applying the same process to subsequent iterates, the update rule for $\mathbf{z}_i^k$ can then be obtained recursively. We propose that, for any $i \in [n]$,

$$\mathbf{z}_i^k - \mathbf{z}^* = [\mathbf{I} - \alpha \sum_{j=1}^{i} \left(\widetilde{\prod}_{t=j+1}^{i} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)}](\mathbf{z}_0^k - \mathbf{z}^*) \tag{C.4}$$

$$- \alpha \sum_{j=1}^{i} \left(\widetilde{\prod}_{t=j+1}^{i} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \omega_{\tau_k(j)}(\mathbf{z}^*).$$

We clarify that the matrix products in (C.4) are in reverse order, and hence reduce to the empty product, which is defined to be $\mathbf{I}$, when $j = i$.

As verified above, (C.4) is satisfied for $i = 1, 2$. We now prove that it holds for any $i \in [n]$ by induction.

Assume (C.4) is true for some $i \in [n]$. Using (C.3) to derive the update of $\mathbf{z}_{i+1}^k$ yields,

$$\mathbf{z}_{i+1}^k - \mathbf{z}^* = \mathbf{z}_i^k - \mathbf{z}^* - \alpha \omega_{\tau_k(i+1)}(\mathbf{z}^*) - \alpha \mathbf{M}_{\tau_k(i+1)}(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha \mathbf{J}_{\tau_k(i+1)}(\mathbf{z}_i^k - \mathbf{z}_0^k)$$

$$= (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(i+1)})(\mathbf{z}_i^k - \mathbf{z}^*) - \alpha(\mathbf{M}_{\tau_k(i+1)} - \mathbf{J}_{\tau_k(i+1)})(\mathbf{z}_0^k - \mathbf{z}^*) - \alpha \omega_{\tau_k(i+1)}(\mathbf{z}^*)$$

$$= (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(i+1)})[\mathbf{I} - \alpha \sum_{j=1}^{i} \left(\widetilde{\prod}_{t=j+1}^{i} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)}](\mathbf{z}_0^k - \mathbf{z}^*)$$

$$- \alpha[\omega_{\tau_k(i+1)}(\mathbf{z}^*) + \sum_{j=1}^{i} (I - \alpha \mathbf{J}_{\tau_k(i+1)}) \left(\widetilde{\prod}_{t=j+1}^{i} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \omega_{\tau_k(j)}(\mathbf{z}^*)]$$

$$- \alpha(\mathbf{M}_{\tau_k(i+1)} - \mathbf{J}_{\tau_k(i+1)})(\mathbf{z}_0^k - \mathbf{z}^*)$$

$$= [\mathbf{I} - \alpha \mathbf{M}_{\tau_k(i+1)} - \alpha \sum_{j=1}^{i} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(i+1)}) \left(\widetilde{\prod}_{t=j+1}^{i} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)}](\mathbf{z}_0^k - \mathbf{z}^*)$$

$$- \alpha \sum_{j=1}^{i+1} \left(\widetilde{\prod}_{t=j+1}^{i+1} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \omega_{\tau_k(j)}(\mathbf{z}^*).$$

It follows that,

$$\mathbf{z}_{i+1}^k - \mathbf{z}^* = [\mathbf{I} - \alpha \sum_{j=1}^{i+1} \left(\widetilde{\prod}_{t=j+1}^{i+1} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)}](\mathbf{z}_0^k - \mathbf{z}^*)$$

$$- \alpha \sum_{j=1}^{i+1} \left(\widetilde{\prod}_{t=j+1}^{i+1} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \omega_{\tau_k(j)}(\mathbf{z}^*),$$

which is simply (C.4) for the iterate $\mathbf{z}_{i+1}^k$. Hence, by induction, (C.4) holds for any $i \in [n]$. As before, the reverse matrix products for the summand $j = i + 1$ is empty (in both the sums) and hence, is equal to $\mathbf{I}$.

Since (C.4) holds for any $i \in [n]$, it holds for $\mathbf{z}_0^{k+1} = \mathbf{z}_n^k$ which gives us the following epoch-level update rule

$$\mathbf{z}_0^{k+1} - \mathbf{z}^* = [\mathbf{I} - \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)}](\mathbf{z}_0^k - \mathbf{z}^*)$$

$$- \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \omega_{\tau_k(j)}(\mathbf{z}^*). \tag{C.5}$$

We simplify the second term in the right hand side of (C.5) using the summation by parts identity. To this end, we define a_j and b_j as

$$a_j = \widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}),$$

$$b_j = \omega_{\tau_k(j)}(\mathbf{z}^*),$$

22

and also note that,

$$\sum_{j=1}^{n} b_j = \sum_{j=1}^{n} \omega_{\tau_k(j)}(\mathbf{z}^*) = n\nu(\mathbf{z}^*) = 0,$$

since $\mathbf{z}^*$ is the unique root of ν. We now apply the summation by parts identity as follows:

$$
\begin{aligned}
\sum_{j=1}^{n} a_j b_j &= a_n \sum_{j=1}^{n} b_j - \sum_{i=1}^{n-1} (a_{i+1} - a_i) \sum_{j=1}^{i} b_j \\
&= -\sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) - \widetilde{\prod}_{t=i+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right] \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*) \\
&= -\alpha \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right] \mathbf{J}_{\tau_k(i+1)} \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*).
\end{aligned}
$$

Hence, we conclude that

$$\sum_{j=1}^{n} \left[\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right] \omega_{\tau_k(j)}(\mathbf{z}^*) = -\alpha \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right] \mathbf{J}_{\tau_k(i+1)} \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*),$$

and introduce $\mathbf{H}_k$ and $\mathbf{r}_k$ as

$$\mathbf{H}_k = \mathbf{I} - \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)},$$

$$\mathbf{r}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right] \mathbf{J}_{\tau_k(i+1)} \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*).$$

By substituting the above expressions in (C.5), we obtain the desired epoch-level update rule for GDA-RR and GDA-SO:

$$\mathbf{z}_0^{k+1} - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^k - \mathbf{z}^*) + \alpha^2 \mathbf{r}_k. \tag{C.6}$$

We observe that the linearized epoch-level update rule (C.6) closely resembles the linearized update rule (C.1) of full-batch GDA as derived in Theorem C.1 with an additional noise term $\alpha^2 \mathbf{r}_k$. In fact, for $n = 1$, which corresponds to the full-batch regime, it is easy to see that $\mathbf{r}_k = 0$ and (C.6) exactly recovers the linearized full-batch GDA update rule (C.1). As we shall see in Theorem C.3, the epoch iterates of GDA-AS also follow the exact same update. The derivation of such an epoch-level update rule which simultaneously handles GDA-RR, GDA-SO and GDA-AS is a cornerstone of our unified analysis. Intuition behind this strategy dates back to the foundational works of Nedic and Bertsekas [37] and Gürbüzbalaban et al. [21], where the authors interpret GD with RR as a stochastic approximation of full-batch GDA, and then obtain asymptotic $O(1/K^2)$ rates using Chung's Lemma. Our linearization strategy (which we also use for analyzing PPM and AGDA) essentially formalizes those insights for minimax optimization in a manner that enables the derivation of non-asymptotic rates.

Similar to the procedure followed in Theorem C.1, we now upper bound $|\mathbf{H}_k|$ as follows:

Step 2: Upper bounding $|\mathbf{H}_k|$ We define the matrix $\mathbf{M}$ as,

$$\mathbf{M} = 1/n \sum_{j=1}^{n} \mathbf{M}_{\tau_k(j)} = 1/n \sum_{j=1}^{n} \int_0^1 \nabla \omega_{\tau_k(j)}(t\mathbf{z}_0^k + (1-t)\mathbf{z}^*) \mathrm{d}t = \int_0^1 \nabla \nu(t\mathbf{z}_0^k + (1-t)\mathbf{z}^*) \mathrm{d}t.$$

Following the same arguments as in Theorem C.1, we note that the l-smoothness and μ-strong monotonicity of ν implies $|\mathbf{M}| \leq l$ and $\mathbf{v}^T \mathbf{M} \mathbf{v} \geq \mu |\mathbf{v}|^2 \; \forall \mathbf{v} \in \mathbb{R}^d$.

We now simplify $\mathbf{H}_k$ by unrolling the sum, beginning with the last index $j = n$:

$$
\begin{aligned}
\mathbf{H}_k &= \mathbf{I} - \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(j)} \\
&= \mathbf{I} - \alpha \mathbf{M}_{\tau_k(n)} - \alpha(\mathbf{I} - \alpha \mathbf{J}_{\tau_k(n)}) \mathbf{M}_{\tau_k(n-1)} - \alpha(\mathbf{I} - \alpha \mathbf{J}_{\tau_k(n)})(\mathbf{I} - \alpha \mathbf{J}_{\tau_k(n-1)}) \mathbf{M}_{\tau_k(n-2)} \\
&\quad - \ldots - \alpha \left(\widetilde{\prod}_{t=2}^{n} (\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}) \right) \mathbf{M}_{\tau_k(1)}.
\end{aligned}
$$

23

Expanding each product and grouping by powers of α, we obtain the following:

$$\mathbf{H}_k = \mathbf{I} - \alpha \sum_{t_1=1}^{n} \mathbf{M}_{\tau_k(t_1)} + \alpha^2 \sum_{1 \le t_1 < t_2 \le n} \mathbf{J}_{\tau_k(t_2)} \mathbf{M}_{\tau_k(t_1)} - \alpha^3 \sum_{1 \le t_1 < t_2 < t_3 \le n} \mathbf{J}_{\tau_k(t_3)} \mathbf{J}_{\tau_k(t_2)} \mathbf{M}_{\tau_k(t_1)}$$
$$+ \ldots + (-1)^n \alpha^n \sum_{1 \le t_1 < \ldots < t_n \le n} \mathbf{J}_{\tau_k(t_n)} \mathbf{J}_{\tau_k(t_{n-1})} \ldots \mathbf{J}_{\tau_k(t_2)} \mathbf{M}_{\tau_k(t_1)}.$$

We note that by definition of $\mathbf{M}$, $\sum_{t_1=1}^{n} \mathbf{M}_{\tau_k(t_1)} = n\mathbf{M}$. Regrouping the remaining terms, we obtain:

$$\mathbf{H}_k = \mathbf{I} - n\alpha\mathbf{M} + \sum_{j=2}^{n} (-1)^j \alpha^j \sum_{1 \le t_1 < t_2 < \ldots < t_j \le n} \mathbf{J}_{\tau_k(t_j)} \mathbf{J}_{\tau_k(t_{j-1})} \ldots \mathbf{J}_{\tau_k(t_2)} \mathbf{M}_{\tau_k(t_1)}.$$

Hence, by the triangle inequality, it can be inferred that

$$|\mathbf{H}_k| \le |\mathbf{I} - n\alpha\mathbf{M}| + \sum_{j=2}^{n} \alpha^j \sum_{1 \le t_1 < t_2 < \ldots < t_j \le n} \left|\mathbf{J}_{\tau_k(t_j)}\right| \left|\mathbf{J}_{\tau_k(t_{j-1})}\right| \ldots \left|\mathbf{J}_{\tau_k(t_2)}\right| \left|\mathbf{M}_{\tau_k(t_1)}\right|. \quad \text{(C.7)}$$

We begin by bounding $|\mathbf{I} - n\alpha\mathbf{M}|$ in a procedure similar to that of Theorem C.1. Observe that for any $\mathbf{v} \in \mathbb{R}^d$

$$|(\mathbf{I} - n\alpha\mathbf{M})\mathbf{v}|^2 = |\mathbf{v}|^2 - n\alpha\mathbf{v}^T\mathbf{M}\mathbf{v} - n\alpha\mathbf{v}^T\mathbf{M}\mathbf{v} + n^2 |\mathbf{M}\mathbf{v}|^2$$
$$\le (1 - 2n\alpha\mu + n^2\alpha^2 l^2) |\mathbf{v}|^2,$$

where the last inequality follows from $|\mathbf{M}\mathbf{v}| \le |\mathbf{M}| |\mathbf{v}| \le l |\mathbf{v}|$ and $\mathbf{v}^T\mathbf{M}^T\mathbf{v} = \mathbf{v}^T\mathbf{M}\mathbf{v} \ge \mu |\mathbf{v}|^2$. As the above holds true for any $\mathbf{v} \in \mathbb{R}^d$, it follows that

$$|\mathbf{I} - n\alpha\mathbf{M}| \le \sqrt{1 - 2n\alpha\mu + n^2\alpha^2 l^2} \le \sqrt{1 - \frac{9n\alpha\mu}{5}} \le 1 - \frac{9n\alpha\mu}{10},$$

where we use $\alpha \le \frac{\mu}{5nl^2}$ in the second and last inequalities.

It now remains to bound the right hand side of (C.7). Note that the constraint $1 \le t_1 < t_2 < \ldots < t_j \le n$ implies that there exist only $\binom{n}{j}$ possible choices for the t_i's. Furthermore, as stated in Step 1, the l-smoothness of the components $\omega_i(\mathbf{z})$ implies that $\left|\mathbf{M}_{\tau_k(i)}\right|, \left|\mathbf{J}_{\tau_k(i)}\right| \le l, \forall i \in [n]$. Hence,

$$\sum_{j=2}^{n} \alpha^j \sum_{1 \le t_1 < t_2 < \ldots < t_j \le n} \left|\mathbf{J}_{\tau_k(t_j)}\right| \left|\mathbf{J}_{\tau_k(t_{j-1})}\right| \ldots \left|\mathbf{J}_{\tau_k(t_2)}\right| \left|\mathbf{M}_{\tau_k(t_1)}\right| \le \sum_{j=2}^{n} \alpha^j \binom{n}{j} l^j.$$

Combining the previous inequalities allows us to bound $|\mathbf{H}_k|$ as follows:

$$|\mathbf{H}_k| \le 1 - \frac{9n\alpha\mu}{10} + \sum_{j=2}^{n} (\alpha l)^j \binom{n}{j} \le 1 - \frac{9n\alpha\mu}{10} + \sum_{j=2}^{n} (\alpha n l)^j \le 1 - \frac{9n\alpha\mu}{10} + \frac{\alpha^2 n^2 l^2}{1 - \alpha n l}$$
$$\le 1 - \frac{9n\alpha\mu}{10} + \frac{5\alpha^2 n^2 l^2}{4} \le 1 - \frac{9n\alpha\mu}{10} + \frac{n\alpha\mu}{4} \le 1 - n\alpha\mu/2,$$

where we substitute $\alpha \le \frac{\mu}{5nl^2} \le 1/5$ wherever required.

We highlight that our analysis so far holds for any permutation $\tau_k \in \mathbb{S}_n$. Consequently, Steps 1 and 2 directly generalize to the adversarial shuffling regime without any modifications. We use this as a starting point for our analysis of GDA-AS in Theorem C.3.

We now proceed to control the magnitude of the noise term $\mathbf{r}_k$ by bounding its expected squared norm $\mathbb{E}[|\mathbf{r}_k|^2]$, where the expectation is taken over the single uniform permutation τ for GDA-SO and over independent uniform permutations $\tau_1, \tau_2, \ldots, \tau_K$ for GDA-RR.

24

Step 3: Upper bounding $\mathbb{E}[|\mathbf{r}_k|^2]$ Recall that, as per Step 1, $\left|\mathbf{J}_{\tau_k(t)}\right| \leq l$. Hence, applying the triangle inequality yields,

$$|\mathbf{r}_k| \leq \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n} |\mathbf{I} - \alpha \mathbf{J}_{\tau_k(t)}| \right] |\mathbf{J}_{\tau_k(i+1)}| \, | \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|$$

$$\leq l(1+\alpha l)^n \sum_{i=1}^{n-1} | \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|$$

$$\leq l(1+\frac{\mu}{5nl})^n \sum_{i=1}^{n-1} | \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)| \leq l e^{1/5} \sum_{i=1}^{n-1} | \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|.$$

To bound $|\mathbf{r}_k|^2$, we use the inequality $(\sum_{i=1}^{n} x_i)^2 \leq n \sum_{i=1}^{n} x_i^2$ which follows from Young's inequality. This yields:

$$|\mathbf{r}_k|^2 \leq l^2 e^{2/5} \left(\sum_{i=1}^{n-1} | \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)| \right)^2 \leq l^2 e^{2/5} \left(\sum_{i=1}^{n-1} i |1/i \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)| \right)^2$$

$$\leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 |1/i \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|^2.$$

We now convert the above inequality into an upper bound for $\mathbb{E}[|\mathbf{r}_k|^2]$. For the sake of clarity, we illustrate the case of GDA-RR and GDA-SO separately. However, we remark that the proof for both cases is practically identical and relies on an application of Lemma B.3 to the term $|1/i \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|^2$.

Step 3A: $\mathbb{E}[|\mathbf{r}_k|^2]$ **for GDA-RR** We write $\mathbb{E}[|\mathbf{r}_k|^2] = \mathbb{E}_{\tau_1,\ldots,\tau_K}[|\mathbf{r}_k|^2]$ to explicitly denote the dependence of the expectation on the uniform random permutations $\tau_1, \ldots, \tau_K$. We then use Lemma B.3 to bound $\mathbb{E}[|\mathbf{r}_k|^2]$ as follows:

$$\mathbb{E}[|\mathbf{r}_k|^2] \leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \mathbb{E}_{\tau_1,\ldots,\tau_K}[|1/i \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|^2]$$

$$\leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \mathbb{E}_{\tau_k}[|1/i \sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|^2]$$

$$\leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \frac{n-i}{i(n-1)} \sigma_*^2 = l^2 e^{2/5} \sigma_*^2 \sum_{i=1}^{n-1} i(n-i)$$

$$\leq l^2 n^3 \sigma_*^2 / 4,$$

where σ_*^2 denotes the gradient variance at the minimax point $\mathbf{z}^*$, defined as $\sigma_*^2 = 1/n \sum_{i=1}^{n} |\omega_i(\mathbf{z}^*)|^2$. We highlight that the second inequality follows from the independence of the uniformly sampled permutations $\tau_1, \ldots, \tau_K$ and the third inequality follows from Lemma B.3 (recall that $1/n \sum_{j=1}^{n} \omega(\mathbf{z}^*) = \nu(\mathbf{z}^*) = 0$).

Step 3B: $\mathbb{E}[|\mathbf{r}_k|^2]$ **for GDA-SO** For GDA-SO, $\tau_1 = \ldots = \tau_K = \tau$ where τ is a uniformly sampled random permutation. Hence, this yields the following:

$$|\mathbf{r}_k|^2 \leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 |1/i \sum_{j=1}^{i} \omega_{\tau(j)}(\mathbf{z}^*)|^2.$$

25

Proceeding along the same lines as Step 3A, we obtain

$$\mathbb{E}[|\mathbf{r}_k|^2] \le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \mathbb{E}_\tau[|1/i \sum_{j=1}^{i} \omega_{\tau(j)}(\mathbf{z}^*)|^2]$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \frac{n-i}{i(n-1)} \sigma_*^2$$

$$\le l^2 n^3 \sigma_*^2/4.$$

Hence, for both GDA-RR and GDA-SO, $\mathbf{r}_k$ is bounded in expectation as $\mathbb{E}[|\mathbf{r}_k|^2] \le l^2 n^3 \sigma_*^2/4$.

Having bounded both $|\mathbf{H}_k|$ and $\mathbb{E}[|\mathbf{r}_k|^2]$, we now proceed to present a complete unified proof of convergence for GDA-RR and GDA-SO.

Step 4: Convergence analysis   Unrolling (C.6) for K epochs and setting $\mathbf{z}_0^1 = \mathbf{z}_0$ gives us the following:

$$\mathbf{z}_0^{K+1} - \mathbf{z}^* = \left[\widetilde{\prod}_{k=1}^{K} \mathbf{H}_k \right] (\mathbf{z}_0 - \mathbf{z}^*) + \alpha^2 \sum_{k=1}^{K} \left[\widetilde{\prod}_{j=k+1}^{K} \mathbf{H}_j \right] \mathbf{r}_k.$$

Using the triangle inequality and the bound $|\mathbf{H}_k| \le 1 - n\alpha\mu/2$ results in:

$$\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right| \le \left[\widetilde{\prod}_{k=1}^{K} |\mathbf{H}_k| \right] |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^{K} \left[\widetilde{\prod}_{j=k+1}^{K} |\mathbf{H}_j| \right] |\mathbf{r}_k|$$

$$\le (1 - n\alpha\mu/2)^K |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{K-k} |\mathbf{r}_k|.$$

By applying Young's inequality, we conclude the following:

$$\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2 \le 2(1 - n\alpha\mu/2)^{2K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} |\mathbf{r}_k|^2$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} |\mathbf{r}_k|^2.$$

Taking expectations (with respect to the uniform random permutations $\tau_1, \ldots, \tau_K$ for RR and τ for SO respectively) on both sides of the above inequality and substituting the upper bound on $\mathbb{E}[|\mathbf{r}_k|^2]$ as derived in Step 3 gives us the following:

$$\mathbb{E}[\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2] \le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} \mathbb{E}[|\mathbf{r}_k|^2]$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + \left(l^2 n^3 \sigma_*^2/2 \right) \alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)}.$$

By simplifying the last term, we obtain the following guarantee that holds for GDA-RR and GDA-SO whenever $\alpha \le \frac{\mu}{5nl^2}$,

$$\mathbb{E}[\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2] \le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{l^2 \sigma_*^2 \alpha^3 n^2 K}{\mu}. \tag{C.8}$$

To complete the proof, we substitute $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)| n^{1/2} K/\mu)/\mu nK\}$ in (C.8).

Under this choice of step-size, using $\alpha \le \frac{2\log(|\nu(\mathbf{z}_0)| n^{1/2} K/\mu)}{\mu nK}$, we bound the second term in the right hand side of (C.8) as

$$\frac{l^2 \sigma_*^2 \alpha^3 n^2 K}{\mu} \le \frac{8l^2 \sigma_*^2}{\mu^4} \frac{\log^3\left(|\nu(\mathbf{z}_0)| n^{1/2} K/\mu \right)}{nK^2}.$$

26

By substituting the above expression into (C.8), we obtain the following:

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq 2e^{-n\alpha\mu K}\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 + \frac{8\kappa^2\sigma_*^2}{\mu^2}\frac{\log^3\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{nK^2}. \tag{C.9}$$

We now consider the following cases:

Case 1: $\frac{\mu}{5nl^2} \leq \frac{2\log\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{\mu nK}$ By our choice of the step-size, $\alpha = \frac{\mu}{5nl^2}$, which implies that

$$2e^{-n\alpha\mu K}\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 = 2e^{-\frac{K}{5\kappa^2}}\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2.$$

By substituting into (C.9), we obtain the following rate for Case 1,

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq 2e^{-\frac{K}{5\kappa^2}}\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 + \frac{8\kappa^2\sigma_*^2}{\mu^2}\frac{\log^3\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{nK^2}. \tag{C.10}$$

Case 2: $\frac{2\log\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{\mu nK} \leq \frac{\mu}{5nl^2}$ By our choice of the step-size, $\alpha = \frac{2\log\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{\mu nK}$. It follows that:

$$2e^{-n\alpha\mu K}\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 = 2e^{-2\log\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)} \leq \frac{\mu\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2}{\left|\nu(\mathbf{z}_0)\right|^2}\frac{2}{nK^2} \leq \frac{2}{nK^2},$$

where the last inequality follows from the μ-strong monotonicity of ν.

Substituting the above inequality into (C.9), we obtain the following rate for Case 2,

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq \frac{2\mu^2 + 8\kappa^2\sigma_*^2\log^3\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{\mu^2 nK^2}. \tag{C.11}$$

Taking the maximum of the right hand side of (C.10) and (C.11), we finally obtain the desired last-iterate convergence rate which holds for both cases:

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq 2e^{-\frac{K}{5\kappa^2}}\left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 + \frac{2\mu^2 + 8\kappa^2\sigma_*^2\log^3\left(\left|\nu(\mathbf{z}_0)\right|n^{1/2}K/\mu\right)}{\mu^2 nK^2}. \tag{C.12}$$

Suppressing constant terms and logarithmic factors, we obtain the following:

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] = \tilde{O}(e^{-K/5\kappa^2} + {}^1\!/nK^2). \tag{C.13}$$

We highlight that the obtained last-iterate convergence guarantee holds for both GDA-RR and GDA-SO, and is applicable for any number of epochs $K \geq 1$. $\square$

Recovering the rates of full-batch GDA   We note that when $n = 1$, i.e., in the non-stochastic regime, GDA-RR/SO reduces to full-batch GDA and $\sigma_*^2 = 0$. Substituting this into (C.12), we note that Theorem C.2 implies a last-iterate convergence guarantee of $O(\exp(-K/5\kappa^2))$, which matches the tight convergence rate of full-batch GDA as established in Theorem C.1. Thus, our analysis of GDA-RR/SO automatically recovers the rates of full-batch GDA and the exponential dependence of $-K/5\kappa^2$ is optimal up to constant factors. We highlight that the above argument holds even when $n > 1$ as long as $\sigma_*^2 = 0$ (which occurs whenever $\mathbf{z}^*$ is the common root of all the operators ω_i).

Comparison with lower bounds and uniform sampling   As discussed earlier in Section 3.2, it is easy to see that the obtained convergence rate for GDA-RR/SO is nearly tight, i.e., differs from the lower bound only by an exponentially decaying term $\exp(-K/5\kappa^2)$. In fact, Theorem C.2 implies an $\tilde{O}(1/nK^2)$ convergence rate (suppressing constants such as κ, μ, σ_* as well as logarithmic factors) when $\exp(-K/5\kappa^2) + 1/nK^2 = O(1/nK^2)$, which holds when K satisfies an epoch requirement of the form $K \geq 10\kappa^2\log(n^{1/2}K)$. Thus, as per Theorem C.2, GDA-RR/SO improves upon SGDA with replacement and matches the lower bound (modulo logarithmic factors) when $K \geq 10\kappa^2\log(n^{1/2}K)$. Furthermore, since our analysis of GDA-RR/SO recovers the rates of full-batch GDA, one can infer that the dependence of this epoch requirement on κ cannot be improved for constant step-sizes (i.e., using constant step-sizes, one cannot obtain $\tilde{O}(1/nK^2)$ rates by imposing an epoch requirement of the form $K \geq C\kappa^a\log(n^{1/2}K)$ with $a < 2$), as it would contradict the optimality of the $-K/5\kappa^2$ exponential dependence. We remark that it might be possible to relax or even remove this epoch requirement for matching the lower bound, by either assuming strong monotonicity of the components ω_i or by using time-varying step-sizes, as was done in Mishchenko et al. [34] and Ahn et al. [1] respectively, in the minimization setting.

27

C.3 Analysis of GDA-AS

Theorem C.3 (Convergence of GDA-AS). *Consider Problem (4) for the μ-strongly monotone operator $\nu(\mathbf{z}) = 1/n \sum_{i=1}^{n} \omega_i(\mathbf{z})$ where each ω_i is l-Lipschitz, but not necessarily monotone. Let $\mathbf{z}^*$ denote the unique root of ν. Then, for any $\alpha \leq \frac{\mu}{5nl^2}$ and $K \geq 1$, the iterates of GDA-AS satisfy the following:*

$$\max_{\tau_1,\ldots,\tau_K \in \mathbb{S}_n} \left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 \leq 2e^{-n\alpha\mu K} \left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 + \frac{3l^2\sigma_*^2\alpha^3 n^3 K}{\mu}.$$

Setting $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)| K/\mu)/\mu nK\}$ results in the following expected last-iterate convergence guarantee for GDA-AS, which holds for any $K \geq 1$:

$$\max_{\tau_1,\ldots,\tau_K \in \mathbb{S}_n} \left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 \leq 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 24\kappa^2\sigma_*^2 \log^3(|\nu(\mathbf{z}_0)|K/\mu)}{\mu^2 K^2}$$

$$= \tilde{O}(e^{-K/5\kappa^2} + 1/K^2).$$

Proof. The iterate-level update rule of GDA-AS can be expressed as

$$\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_{i-1}^k), \tag{C.14}$$

where $k \in [K]$, $i \in [n]$ and $\tau_k \in \mathbb{S}_n$ denotes an arbitrary permutation chosen by the adversary at the start of epoch k. Recall that Steps 1 and 2 of Theorem C.2 hold for any permutation $\tau_k \in \mathbb{S}_n$, and consequently, they are directly applicable to GDA-AS. As a result, we can conclude that the epoch-level update rule of GDA-AS is the same as that of GDA-RR/SO, and is given by:

$$\mathbf{z}_0^{k+1} - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^k - \mathbf{z}^*) + \alpha^2\mathbf{r}_k. \tag{C.15}$$

Note that $\mathbf{H}_k$ and $\mathbf{r}_k$ are as defined in Theorem C.2, and are given by:

$$\mathbf{H}_k = \mathbf{I} - \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n}(\mathbf{I} - \alpha\mathbf{J}_{\tau_k(t)})\right)\mathbf{M}_{\tau_k(j)},$$

$$\mathbf{r}_k = \sum_{i=1}^{n-1}\left[\widetilde{\prod}_{t=i+2}^{n}(\mathbf{I} - \alpha\mathbf{J}_{\tau_k(t)})\right]\mathbf{J}_{\tau_k(i+1)}\sum_{j=1}^{i}\omega_{\tau_k(j)}(\mathbf{z}^*),$$

where, as before, $\mathbf{M}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_0^k + (1-t)\mathbf{z}^*)dt$ and $\mathbf{J}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_i^k + (1-t)\mathbf{z}_0^k)dt$. Furthermore, by Step 2 of Theorem C.2, $|\mathbf{H}_k| \leq 1 - n\alpha\mu/2$.

Our analysis now proceeds by controlling the influence of the noise term $\mathbf{r}_k$. Recall that this was achieved in our analysis of GDA-RR/SO by using Lemma B.3 to upper bound $\mathbb{E}[|\mathbf{r}_k|^2]$. However, the same strategy does not apply to GDA-AS since the permutations $\tau_1, \ldots, \tau_K$ are chosen by the adversary as per an arbitrary strategy that is unknown to us. We circumvent this by instead deriving an upper bound for $|\mathbf{r}_k|^2$ that holds uniformly for any permutation $\tau_k \in \mathbb{S}_n$.

We begin by repeating some initial steps from Step 3 of Theorem C.2. Applying $\left|\mathbf{J}_{\tau_k(t)}\right| \leq l$ gives us the following:

$$|\mathbf{r}_k| \leq \sum_{i=1}^{n-1}\left[\widetilde{\prod}_{t=i+2}^{n}\left|\mathbf{I} - \alpha\mathbf{J}_{\tau_k(t)}\right|\right]\left|\mathbf{J}_{\tau_k(i+1)}\right|\left|\sum_{j=1}^{i}\omega_{\tau_k(j)}(\mathbf{z}^*)\right| \leq l(1+\alpha l)^n \sum_{i=1}^{n-1}\left|\sum_{j=1}^{i}\omega_{\tau_k(j)}(\mathbf{z}^*)\right|$$

$$\leq l(1+\frac{\mu}{5nl})^n \sum_{i=1}^{n-1}\left|\sum_{j=1}^{i}\omega_{\tau_k(j)}(\mathbf{z}^*)\right| \leq le^{1/5}\sum_{i=1}^{n-1}\left|\sum_{j=1}^{i}\omega_{\tau_k(j)}(\mathbf{z}^*)\right|.$$

28

By Young's inequality, we conclude,

$$|\mathbf{r}_k|^2 \le l^2 e^{2/5} \left(\sum_{i=1}^{n-1} |\sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)| \right)^2$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} |\sum_{j=1}^{i} \omega_{\tau_k(j)}(\mathbf{z}^*)|^2$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{i} |\omega_{\tau_k(j)}(\mathbf{z}^*)|^2$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{n} |\omega_{\tau_k(j)}(\mathbf{z}^*)|^2 = l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{n} |\omega_j(\mathbf{z}^*)|^2$$

$$= l^2 e^{2/5}(n-1)(n\sigma_*^2) \sum_{i=1}^{n-1} i \le 3l^2 n^4 \sigma_*^2/4.$$

We note that the above sequence of inequalities holds for any permutation τ_k, and hence, the upper bound $|\mathbf{r}_k|^2 \le 3l^2 n^4 \sigma_*^2/4$ holds for any $\tau_k \in \mathbb{S}_n$, $k \in [K]$.

Having bounded both $|\mathbf{H}_k|$ and $|\mathbf{r}_k|$, we now proceed to derive the last-iterate convergence guarantee. Unrolling (C.15) for K epochs and setting $\mathbf{z}_0^1 = \mathbf{z}_0$ gives us the following:

$$\mathbf{z}_0^{K+1} - \mathbf{z}^* = \left[\widetilde{\prod}_{k=1}^{K} \mathbf{H}_k \right] (\mathbf{z}_0 - \mathbf{z}^*) + \alpha^2 \sum_{k=1}^{K} \left[\widetilde{\prod}_{j=k+1}^{K} \mathbf{H}_k \right] \mathbf{r}_k.$$

By the triangle inequality and the bound $|\mathbf{H}_k| \le 1 - n\alpha\mu/2$, we infer,

$$\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right| \le \left[\widetilde{\prod}_{k=1}^{K} |\mathbf{H}_k| \right] |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^{K} \left[\widetilde{\prod}_{j=k+1}^{K} |\mathbf{H}_k| \right] |\mathbf{r}_k|$$

$$\le (1 - n\alpha\mu/2)^K |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{K-k} |\mathbf{r}_k|.$$

By applying Young's inequality, we can conclude that,

$$\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2 \le 2(1 - n\alpha\mu/2)^{2K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} |\mathbf{r}_k|^2$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} |\mathbf{r}_k|^2$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + (1.5 l^2 n^4 \sigma_*^2) \alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)}$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{3l^2 \sigma_*^2 \alpha^3 n^3 K}{\mu}.$$

Since the above inequality holds for any sequence of permutations $\tau_1, \dots, \tau_K$ chosen by the adversary, we obtain the following uniform guarantee for GDA-AS, whenever $\alpha \le \frac{\mu}{5nl^2}$,

$$\max_{\tau_1, \dots, \tau_K \in \mathbb{S}_n} \left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2 \le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{3l^2 \sigma_*^2 \alpha^3 n^3 K}{\mu}. \tag{C.16}$$

To complete the proof, we substitute $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)| K/\mu)/\mu n K\}$ in (C.16) and proceed in a manner similar to Theorem C.2.

29

Under this choice of step-size, using $\alpha \le \frac{2\log(|\nu(\mathbf{z}_0)|K/\mu)}{\mu n K}$, we bound the second term in the right hand side of (C.16) as

$$\frac{3l^2\sigma_*^2\alpha^3 n^3 K}{2\mu} \le \frac{24l^2\sigma_*^2}{\mu^4}\frac{\log^3\left(|\nu(\mathbf{z}_0)|\,K/\mu\right)}{K^2}.$$

Substituting the above inequality into (C.16), we obtain,

$$\max_{\tau_1,\ldots,\tau_K\in\mathbb{S}_n}\left|\mathbf{z}_0^{K+1}-\mathbf{z}^*\right|^2 \le 2e^{-n\alpha\mu K}\left|\mathbf{z}_0-\mathbf{z}^*\right|^2 + \frac{24\kappa^2\sigma_*^2}{\mu^2}\frac{\log^3\left(|\nu(\mathbf{z}_0)|\,K/\mu\right)}{K^2}. \qquad \text{(C.17)}$$

We now consider the following cases:

Case 1: $\frac{\mu}{5nl^2} \le \frac{2\log(|\nu(\mathbf{z}_0)|K/\mu)}{\mu n K}$ By our choice of the step-size, $\alpha = \frac{\mu}{5nl^2}$. It follows that:

$$2e^{-n\alpha\mu K}\left|\mathbf{z}_0-\mathbf{z}^*\right|^2 = 2e^{-K/(5\kappa^2)}\left|\mathbf{z}_0-\mathbf{z}^*\right|^2.$$

Substituting into (C.17), we obtain the following rate for Case 1,

$$\max_{\tau_1,\ldots,\tau_K\in\mathbb{S}_n}\left|\mathbf{z}_0^{K+1}-\mathbf{z}^*\right|^2 \le 2e^{\frac{-K}{5\kappa^2}}\left|\mathbf{z}_0-\mathbf{z}^*\right|^2 + \frac{24\kappa^2\sigma_*^2}{\mu^2}\frac{\log^3\left(|\nu(\mathbf{z}_0)|\,K/\mu\right)}{K^2}. \qquad \text{(C.18)}$$

Case 2: $\frac{2\log(|\nu(\mathbf{z}_0)|K/\mu)}{\mu n K} \le \frac{\mu}{5nl^2}$ By our choice of the step-size, $\alpha = \frac{2\log(|\nu(\mathbf{z}_0)|K/\mu)}{\mu n K}$. It follows that:

$$2e^{-n\alpha\mu K}\left|\mathbf{z}_0-\mathbf{z}^*\right|^2 = 2e^{-2\log(|\nu(\mathbf{z}_0)|K/\mu)} \le \frac{\mu\left|\mathbf{z}_0-\mathbf{z}^*\right|^2}{|\nu(\mathbf{z}_0)|^2}\frac{2}{K^2} \le \frac{2}{K^2},$$

where the last inequality follows from the μ-strong monotonicity of ν.

Substituting the above inequality into (C.17), we obtain the following rate for Case 2,

$$\max_{\tau_1,\ldots,\tau_K\in\mathbb{S}_n}\left|\mathbf{z}_0^{K+1}-\mathbf{z}^*\right|^2 \le \frac{2\mu^2+24\kappa^2\sigma_*^2\log^3(|\nu(\mathbf{z}_0)|K/\mu)}{\mu^2 K^2}. \qquad \text{(C.19)}$$

By taking the maximum of the right hand side of (C.18) and (C.19), we finally obtain the desired last-iterate convergence rate.

$$\max_{\tau_1,\ldots,\tau_K\in\mathbb{S}_n}\left|\mathbf{z}_0^{K+1}-\mathbf{z}^*\right|^2 \le 2e^{-K/5\kappa^2}|\mathbf{z}_0-\mathbf{z}^*|^2 + \frac{2\mu^2+24\kappa^2\sigma_*^2\log^3(|\nu(\mathbf{z}_0)|K/\mu)}{\mu^2 K^2}. \qquad \text{(C.20)}$$

Suppressing constant terms and logarithmic factors, we get,

$$\max_{\tau_1,\ldots,\tau_K\in\mathbb{S}_n}\left|\mathbf{z}_0^{K+1}-\mathbf{z}^*\right|^2 = \tilde{O}(e^{-K/5\kappa^2} + {}^1\!/_{K^2}). \qquad \text{(C.21)}$$

$\square$

Comparison with lower bounds   By repeating the same arguments as those used for GDA-RR/SO, it is easy to see that our analysis of GDA-AS recovers the tight rates of full-batch GDA in the deterministic, i.e., $n = 1$ regime. Furthermore, Theorem C.2 implies an $\tilde{O}(1/K^2)$ convergence rate, which matches the lower bound modulo logarithmic factors, when $e^{-K/5\kappa^2} + {}^1\!/_{K^2} = O(1/K^2)$, i.e., when K satisfies an epoch requirement of the form $K \ge 10\kappa^2\log(K)$. As before, since our analysis of GDA-AS recovers the rates of full-batch GDA, the κ^2 dependence of the epoch requirement cannot be improved for constant step-sizes.

D   Analysis of PPM without Replacement

We now present the convergence proofs of the proximal point method without replacement (i.e. PPM-RR/SO/AS). The structure of these results largely mirror that of our analysis of GDA without replacement, since the key techniques used in our analysis is an adaptation of the linearization technique to PPM and the application of Lemma B.3 to control the influence of noise in PPM without replacement. To this end, we begin by presenting a linearization-based proof of convergence of full batch PPM, followed by a unified proof of convergence of PPM-RR and PPM-SO, and conclude with the extension of our proof techniques to the adversarial shuffling regime by analyzing PPM-AS.

30

D.1 Analysis of Full-batch PPM by Linearization

Theorem D.1 (Convergence of full-batch PPM). *Consider the l-Lipschitz and μ-strongly monotone operator $\nu : \mathbb{R}^d \to \mathbb{R}^d$ and let $\mathbf{z}^*$ denote the unique root of ν. For any step-size $0 < \alpha < 1/l$, the iterates $\mathbf{z}_k$ of full-batch PPM satisfy the following recurrence:*

$$\mathbf{z}_{k+1} - \mathbf{z}^* = (\mathbf{I} + \alpha \mathbf{M}_{k+1})^{-1}(\mathbf{z}_k - \mathbf{z}^*),$$

where $\mathbf{I} + \alpha \mathbf{M}_{k+1}$ is invertible and its singular values are lower bounded as $\sigma_{\min}(\mathbf{I} + \alpha \mathbf{M}_{k+1}) \geq 1 + \alpha\mu$. Consequently, setting $\alpha = 1/2l$ gives us the following last iterate convergence guarantee:

$$|\mathbf{z}_{K+1} - \mathbf{z}^*|^2 \leq (1 + 1/2\kappa)^{-2K} |\mathbf{z}_0 - \mathbf{z}^*|^2.$$

Proof. The iterates of full-batch PPM with step-size α satisfy

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \alpha\nu(\mathbf{z}_{k+1}).$$

From the above equation, we conclude that $\mathbf{z}_{k+1}$ is a fixed point of the operator $\zeta(\mathbf{z}) = \mathbf{z}_k - \alpha\nu(\mathbf{z})$. Since $\alpha < 1/l$, ζ is a contraction mapping. Hence, by the Banach Fixed Point Theorem, we conclude that $\mathbf{z}_{k+1}$ exists and is uniquely defined. This allows us to proceed in a manner similar to Theorem C.1 and linearize $\nu(\mathbf{z}_{k+1})$ about $\mathbf{z}^*$. Thus, by the Lipschitz continuity of ν and the Fundamental Theorem of Calculus for Lebesgue Integrals:

$$\nu(\mathbf{z}_{k+1}) = \nu(\mathbf{z}^*) + \int_0^1 \nabla\nu(t\mathbf{z}_{k+1} + (1-t)\mathbf{z}^*)(\mathbf{z}_{k+1} - \mathbf{z}^*)\mathrm{d}t$$
$$= \mathbf{M}_{k+1}(\mathbf{z}_{k+1} - \mathbf{z}^*),$$

where $\mathbf{M}_{k+1} = \int_0^1 \nabla\nu(t\mathbf{z}_{k+1} + (1-t)\mathbf{z}^*)\mathrm{d}t$ is well defined by Lemma B.1. Substituting this expansion into the PPM update rule gives us the following:

$$\mathbf{z}_k - \mathbf{z}^* = (\mathbf{I} + \alpha \mathbf{M}_{k+1})(\mathbf{z}_{k+1} - \mathbf{z}^*). \tag{D.1}$$

Using the same arguments as in Theorem C.1, we conclude that the μ-strong monotonicity of ν implies $\mathbf{v}^T \mathbf{M}_{k+1}\mathbf{v} \geq \mu |\mathbf{v}|^2 \ \forall \mathbf{v} \in \mathbb{R}^d$. Using this, we lower bound the singular values of $\mathbf{I} + \alpha \mathbf{M}_{k+1}$ as follows:

Note that for any $\mathbf{v} \in \mathbb{R}^d$,

$$|(\mathbf{I} + \alpha \mathbf{M}_{k+1})\mathbf{v}|^2 = |\mathbf{v}|^2 + \alpha\mathbf{v}^T \mathbf{M}_{k+1}\mathbf{v} + \alpha\mathbf{v}^T \mathbf{M}_{k+1}^T \mathbf{v} + \alpha^2 |\mathbf{M}_{k+1}\mathbf{v}|^2$$
$$\geq (1 + 2\alpha\mu + \alpha^2\mu^2) |\mathbf{v}|^2,$$

where the last inequality follows from the Cauchy-Schwarz inequality since $|\mathbf{v}| |\mathbf{M}_{k+1}\mathbf{v}| \geq \mathbf{v}^T \mathbf{M}_{k+1}\mathbf{v} \geq \mu |\mathbf{v}|^2$.

Thus, the singular values of $\mathbf{I} + \alpha \mathbf{M}_{k+1}$ are lower bounded as $\sigma_{\min}(\mathbf{I} + \alpha \mathbf{M}_{k+1}) \geq 1 + \alpha\mu > 0$. Since $\mathbf{I} + \mathbf{M}_{k+1}$ is a square matrix, this implies that $\mathbf{I} + \alpha \mathbf{M}_{k+1}$ is invertible and

$$\left|(\mathbf{I} + \alpha \mathbf{M}_{k+1})^{-1}\right| = \sigma_{\max}((\mathbf{I} + \alpha \mathbf{M}_{k+1})^{-1}) = 1/\sigma_{\min}(\mathbf{I} + \alpha \mathbf{M}_{k+1}) \leq (1 + \alpha\mu)^{-1}.$$

It follows that,

$$\mathbf{z}_{k+1} - \mathbf{z}^* = (\mathbf{I} + \alpha \mathbf{M}_{k+1})^{-1}(\mathbf{z}_k - \mathbf{z}^*)$$
$$\implies |\mathbf{z}_{k+1} - \mathbf{z}^*|^2 \leq (1 + \alpha\mu)^{-2} |\mathbf{z}_k - \mathbf{z}^*|^2.$$

Thus, by setting $\alpha = 1/2l$, we obtain the following convergence guarantee

$$|\mathbf{z}_{K+1} - \mathbf{z}^*|^2 \leq (1 + 1/2\kappa)^{-2K} |\mathbf{z}_0 - \mathbf{z}^*|^2.$$

$\square$

31

D.2 A Unified Analysis of PPM-RR and PPM-SO

Theorem D.2 (Convergence of PPM-RR/SO). *Consider Problem* (4) *for the μ-strongly monotone operator $\nu(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\omega_i(\mathbf{z})$ where each ω_i is l-Lipschitz, but not necessarily monotone. Let $\mathbf{z}^*$ denote the unique root of ν. Then, for any $\alpha \leq \frac{\mu}{5nl^2}$ and $K \geq 1$, the iterates of PPM-RR and PPM-SO satisfy the following:*

$$\mathbb{E}[|\mathbf{z}_0^{K+1} - \mathbf{z}^*|^2] \leq 2e^{-n\alpha\mu K}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{l^2\sigma_*^2\alpha^3 n^2 K}{\mu}.$$

Setting $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)|n^{1/2}K/\mu)/\mu nK\}$ results in the following expected last-iterate convergence guarantee for both PPM-RR and PPM-SO, which holds for any $K \geq 1$:

$$\mathbb{E}[|\mathbf{z}_0^{K+1} - \mathbf{z}^*|^2] \leq 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 8\kappa^2\sigma_*^2\log^3(|\nu(\mathbf{z}_0)|n^{1/2}K/\mu)}{\mu^2 nK^2}$$

$$= \tilde{O}(e^{-K/5\kappa^2} + 1/nK^2).$$

Proof. Without loss of generality, we express the iterate-level update rule of PPM-RR and PPM-SO jointly as

$$\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_i^k), \tag{D.2}$$

where $k \in [K]$ and $i \in [n]$. Similar to Theorem D.1, the Banach Fixed Point Theorem guarantees that $\mathbf{z}_i^k$ is well defined and unique since $\alpha \leq \mu/5nl^2 < 1/l$.

We note that $\tau_k \sim \text{Uniform}(\mathbb{S}_n)$ for every $k \in [K]$ for PPM-RR (i.e., τ_k is a permutation of $[n]$ that is resampled uniformly at every epoch) whereas for PPM-SO, $\tau_k = \tau \ \forall k \in [K]$ where $\tau \sim \text{Uniform}(\mathbb{S}_n)$ (i.e., the permutation τ is uniformly sampled before the first epoch, then reused for all subsequent epochs). We also define the transposed permutation $\tilde{\tau}_k$ as $\tilde{\tau}_k(i) = \tau_k(n+1-i)$ for any $i \in [n]$.

The remainder of our proof follows the structure of Theorem C.2. In particular, we apply the linearization technique as illustrated in Theorem D.1 to PPM without replacement and derive a linearized epoch-level update rule of the form $\mathbf{z}_0^k - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^{k+1} - \mathbf{z}^*) + \alpha^2\mathbf{r}_k$. This is followed by lower bounding $\sigma_{\min}(\mathbf{H}_k)$ using the Lipschitz continuity and strong monotonicity of μ and upper bounding $\mathbb{E}[|\mathbf{r}_k|^2]$ using Lemma B.3. The proof is completed by unrolling the update rule for K epochs and carefully choosing the step-size.

Step 1: Linearized epoch-level update rule The approach for this step mirrors that of Theorem C.2 and hinges on the insight that for small enough step-sizes, the dynamics of the iterates $\mathbf{z}_0^{k+1}$ of PPM without replacement can be treated as a noisy version of full-batch PPM. Complementing this insight with the linearization-based analysis of full-batch PPM in Theorem D.1 motivates the following decomposition:

$$\omega_{\tau_k(i)}(\mathbf{z}_i^k) = \omega_{\tau_k(i)}(\mathbf{z}^*) + [\omega_{\tau_k(i)}(\mathbf{z}_n^k) - \omega_{\tau_k(i)}(\mathbf{z}^*)] + [\omega_{\tau_k(i)}(\mathbf{z}_i^k) - \omega_{\tau_k(i)}(\mathbf{z}_n^k)]$$

$$= \omega_{\tau_k(i)}(\mathbf{z}^*) + \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_n^k + (1-t)\mathbf{z}^*)(\mathbf{z}_n^k - \mathbf{z}^*)\mathrm{d}t$$

$$+ \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_i^k + (1-t)\mathbf{z}_n^k)(\mathbf{z}_i^k - \mathbf{z}_n^k)\mathrm{d}t.$$

We further define $\mathbf{M}_{\tau_k(i)}$ and $\mathbf{J}_{\tau_k(i)}$ as follows:

$$\mathbf{M}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_n^k + (1-t)\mathbf{z}^*)\mathrm{d}t,$$

$$\mathbf{J}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_i^k + (1-t)\mathbf{z}_n^k)\mathrm{d}t.$$

By repeating the same arguments as in Theorem D.1, we note that the l-Lipschitz continuity of $\omega_i(\mathbf{z}) \ \forall i \in [n]$ implies that $\mathbf{M}_{\tau_k(i)}$ and $\mathbf{J}_{\tau_k(i)}$ are well defined and bounded as $|\mathbf{M}_{\tau_k(i)}| \leq l$ and $|\mathbf{J}_{\tau_k(i)}| \leq l$. It follows that

$$\omega_{\tau_k(i)}(\mathbf{z}_i^k) = \omega_{\tau_k(i)}(\mathbf{z}^*) + \mathbf{M}_{\tau_k(i)}(\mathbf{z}_n^k - \mathbf{z}^*) + \mathbf{J}_{\tau_k(i)}(\mathbf{z}_i^k - \mathbf{z}_n^k). \tag{D.3}$$

32

Substituting (D.3) into the update equation (C.2) for $\mathbf{z}_n^k$ yields,

$$\mathbf{z}_{n-1}^k - \mathbf{z}^* = (\mathbf{I} + \alpha\mathbf{M}_{\tau_k(n)})(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\mathbf{J}_{\sigma_k(n)}(\mathbf{z}_n^k - \mathbf{z}_n^k) + \alpha\omega_{\tau_k(n)}(\mathbf{z}^*)$$

$$= (\mathbf{I} + \alpha\mathbf{M}_{\tilde{\tau}_k(1)})(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\omega_{\tilde{\tau}_k(1)}(\mathbf{z}^*).$$

Repeating the same for $\mathbf{z}_{n-1}^k$ results in the following:

$$\mathbf{z}_{n-2}^k - \mathbf{z}^* = \mathbf{z}_{n-1}^k - \mathbf{z}^* + \alpha\omega_{\tau_k(n-1)}(\mathbf{z}^*) + \alpha\mathbf{M}_{\tau_k(n-1)}(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\mathbf{J}_{\tau_k(n-1)}(\mathbf{z}_{n-1}^k - \mathbf{z}_n^k)$$

$$= \mathbf{z}_{n-1}^k - \mathbf{z}^* + \alpha\omega_{\tilde{\tau}_k(2)}(\mathbf{z}^*) + \alpha\mathbf{M}_{\tilde{\tau}_k(2)}(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\mathbf{J}_{\tilde{\tau}_k(2)}(\mathbf{z}_{n-1}^k - \mathbf{z}_n^k)$$

$$= (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(2)})(\mathbf{z}_{n-1}^k - \mathbf{z}^*) + \alpha(\mathbf{M}_{\tilde{\tau}_k(2)} - \mathbf{J}_{\tilde{\tau}_k(2)})(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\omega_{\tilde{\tau}_k(2)}(\mathbf{z}^*)$$

$$= (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(2)})[(\mathbf{I} + \alpha\mathbf{M}_{\tilde{\tau}_k(1)})(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\omega_{\tilde{\tau}_k(1)}(\mathbf{z}^*)]$$

$$+ \alpha(\mathbf{M}_{\tilde{\tau}_k(2)} - \mathbf{J}_{\tilde{\tau}_k(2)})(\mathbf{z}_n^k - \mathbf{z}^*) + \alpha\omega_{\tilde{\tau}_k(2)}(\mathbf{z}^*)$$

$$= [(\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(2)})(\mathbf{I} + \alpha\mathbf{M}_{\tilde{\tau}_k(1)}) + \alpha(\mathbf{M}_{\tilde{\tau}_k(2)} - \mathbf{J}_{\tilde{\tau}_k(2)})](\mathbf{z}_n^k - \mathbf{z}^*)$$

$$+ \alpha[\omega_{\tilde{\tau}_k(2)}(\mathbf{z}^*) + (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(2)})\omega_{\tilde{\tau}_k(1)}(\mathbf{z}^*)]$$

$$= [\mathbf{I} + \alpha\mathbf{M}_{\tilde{\tau}_k(2)} + \alpha(\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(2)})\mathbf{M}_{\tilde{\tau}_k(1)}](\mathbf{z}_n^k - \mathbf{z}^*)$$

$$+ \alpha[\omega_{\tilde{\tau}_k(2)}(\mathbf{z}^*) + (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(2)})\omega_{\tilde{\tau}_k(1)}(\mathbf{z}^*)].$$

By applying the same process for preceeding iterates and substituting $\mathbf{z}_n^k = \mathbf{z}_0^{k+1}$, we obtain the following epoch-level update rule for PPM-RR and PPM-SO

$$\mathbf{z}_0^k - \mathbf{z}^* = [\mathbf{I} + \alpha\sum_{j=1}^n \left(\widetilde{\prod}_{t=j+1}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)})\right) \mathbf{M}_{\tilde{\tau}_k(j)}](\mathbf{z}_0^{k+1} - \mathbf{z}^*)$$

$$+ \alpha\sum_{j=1}^n \left(\widetilde{\prod}_{t=j+1}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)})\right) \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*). \tag{D.4}$$

We clarify that the matrix products in (D.4) are in reverse order, and hence reduce to the empty product, which is defined to be $\mathbf{I}$, when $j = i$. Furthermore $\tilde{\tau}_k$ denotes the transposed permutation $\tilde{\tau}_k(i) = \tau_k(n + 1 - i)$.

We simplify the second term in the right hand side of (D.4) using the summation by parts identity. To this end, we define a_j and b_j as

$$a_j = \widetilde{\prod}_{t=j+1}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)}),$$

$$b_j = \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*),$$

and observe that,

$$\sum_{j=1}^n b_j = \sum_{j=1}^n \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*) = n\nu(\mathbf{z}^*) = 0,$$

since $\mathbf{z}^*$ is the unique root of ν. We now apply the summation by parts identity to obtain the following:

$$\sum_{j=1}^n a_j b_j = a_n \sum_{j=1}^n b_j - \sum_{i=1}^{n-1}(a_{i+1} - a_i)\sum_{j=1}^i b_j$$

$$= -\sum_{i=1}^{n-1}\left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)}) - \widetilde{\prod}_{t=i+1}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)})\right]\sum_{j=1}^i \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)$$

$$= \alpha\sum_{i=1}^{n-1}\left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)})\right]\mathbf{J}_{\tilde{\tau}_k(i+1)}\sum_{j=1}^i \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*).$$

Hence, we conclude that:

$$\sum_{j=1}^n\left[\widetilde{\prod}_{t=j+1}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)})\right]\omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*) = \alpha\sum_{i=1}^{n-1}\left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)})\right]\mathbf{J}_{\tilde{\tau}_k(i+1)}\sum_{j=1}^i \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*).$$

33

We define $\mathbf{H}_k$ and $\mathbf{r}_k$ as

$$\mathbf{H}_k = \mathbf{I} + \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} + \alpha \mathbf{J}_{\tilde{\tau}_k(t)}) \right) \mathbf{M}_{\tilde{\tau}_k(j)},$$

$$\mathbf{r}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n} (\mathbf{I} + \alpha \mathbf{J}_{\tilde{\tau}_k(t)}) \right] \mathbf{J}_{\tilde{\tau}_k(i+1)} \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*),$$

and substitute the above expressions in (D.4). This gives us the following noisy linearized update rule for the epoch iterates $\mathbf{z}_0^k$ of PPM-RR and PPM-SO:

$$\mathbf{z}_0^k - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^{k+1} - \mathbf{z}^*) + \alpha^2 \mathbf{r}_k. \tag{D.5}$$

We observe that (D.5) closely resembles the linearized update rule (D.1) of full-batch PPM as derived in Theorem D.1 with an additive noise term $\alpha^2 \mathbf{r}_k$. In fact, for $n = 1$, which corresponds to the full-batch regime, it is easy to see that $\mathbf{r}_k = 0$ and (D.5) reduces to (D.1). As we shall see in Theorem D.3, the same update rule also applies to PPM-AS. Similar to our analysis of GDA, the derivation of an epoch-level update rule that simultaneously handles PPM-RR, PPM-SO and PPM-AS is a key component of our analysis.

Thus, we proceed in a manner similar to Theorem D.1 by lower bounding the minimum singular value of $\mathbf{H}_k$.

Step 2: Lower bounding $\sigma_{\min}(\mathbf{H}_k)$ We define the matrix $\mathbf{M}$ as,

$$\mathbf{M} = 1/n \sum_{j=1}^{n} \mathbf{M}_{\tilde{\tau}_k(j)} = 1/n \sum_{j=1}^{n} \int_0^1 \nabla \omega_{\tilde{\tau}_k(j)}(t\mathbf{z}_n^k + (1-t)\mathbf{z}^*)\mathrm{d}t = \int_0^1 \nabla \nu(t\mathbf{z}_n^k + (1-t)\mathbf{z}^*)\mathrm{d}t.$$

By following the same arguments as Theorem D.1, we note that the l-smoothness and μ-strong monotonicity of ν implies that $|\mathbf{M}| \leq l$ and $\mathbf{v}^T\mathbf{M}\mathbf{v} \geq \mu \, |\mathbf{v}|^2 \; \forall \mathbf{v} \in \mathbb{R}^d$.

By expanding the product terms in $\mathbf{H}_k$, we obtain the following:

$$\mathbf{H}_k = \mathbf{I} + \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n} (\mathbf{I} + \alpha \mathbf{J}_{\tilde{\tau}_k(t)}) \right) \mathbf{M}_{\tilde{\tau}_k(j)}$$

$$= \mathbf{I} + n\alpha\mathbf{M} + \sum_{j=2}^{n} \alpha^j \sum_{1 \leq t_1 < t_2 < \ldots < t_j \leq n} \mathbf{J}_{\tilde{\tau}_k(t_j)}\mathbf{J}_{\tilde{\tau}_k(t_{j-1})} \ldots \mathbf{J}_{\tilde{\tau}_k(t_2)}\mathbf{M}_{\tilde{\tau}_k(t_1)}.$$

We now apply Weyl's inequality for singular value perturbations, which gives us the following bound,

$$\sigma_{\min}(\mathbf{H}_k) \geq \sigma_{\min}(\mathbf{I}+n\alpha\mathbf{M}) - \sum_{j=2}^{n} \alpha^j \sum_{1 \leq t_1 < t_2 < \ldots < t_j \leq n} \left| \mathbf{J}_{\tilde{\tau}_k(t_j)} \right| \left| \mathbf{J}_{\tilde{\tau}_k(t_{j-1})} \right| \ldots \left| \mathbf{J}_{\tilde{\tau}_k(t_2)} \right| \left| \mathbf{M}_{\tilde{\tau}_k(t_1)} \right|. \tag{D.6}$$

Following the same steps as Theorem D.1, we conclude that $\mathbf{v}^T\mathbf{M}\mathbf{v} \geq \mu \, |\mathbf{v}|^2$ implies $\sigma_{\min}(\mathbf{I} + n\alpha\mathbf{M}) \geq 1 + n\alpha\mu$. Furthermore, we observe that,

$$\sum_{j=2}^{n} \alpha^j \sum_{1 \leq t_1 < t_2 < \ldots < t_j \leq n} \left| \mathbf{J}_{\tilde{\tau}_k(t_j)} \right| \left| \mathbf{J}_{\tilde{\tau}_k(t_{j-1})} \right| \ldots \left| \mathbf{J}_{\tilde{\tau}_k(t_2)} \right| \left| \mathbf{M}_{\tilde{\tau}_k(t_1)} \right| \leq \sum_{j=2}^{n} \alpha^j \binom{n}{j} l^j.$$

Substituting the above bound in (D.6) gives us the following:

$$\sigma_{\min}(\mathbf{H}_k) \geq 1 + n\alpha\mu - \sum_{j=2}^{n} (\alpha l)^j \binom{n}{j} \geq 1 + n\alpha\mu - \sum_{j=2}^{n} (\alpha n l)^j \geq 1 + n\alpha\mu - \frac{\alpha^2 n^2 l^2}{1 - \alpha n l}$$

$$\geq 1 + n\alpha\mu - \frac{5\alpha^2 n^2 l^2}{4} \geq 1 + 3n\alpha\mu/4,$$

where we substitute $\alpha \leq \frac{\mu}{5nl^2} \leq 1/5$ wherever required. Furthermore, since $\sigma_{\min}(\mathbf{H}_k) \geq 1 + 3n\alpha\mu/4 > 0$, $\mathbf{H}_k$ is invertible and $\left| \mathbf{H}_k^{-1} \right| \leq (1 + 3n\alpha\mu/4)^{-1} \leq (1 - n\alpha\mu/2)$, which holds since $n\alpha\mu \leq 1/5$.

34

As before, our analysis so far holds for any permutation $\tau_k \in \mathbb{S}_n$. Consequently, Steps 1 and 2 directly generalize to the adversarial shuffling regime without any modifications. We use this as a starting point for our analysis of PPM-AS in Theorem D.3.

We now control the magnitude of the noise term $\mathbf{r}_k$ by bounding $\mathbb{E}[|\mathbf{r}_k|^2]$ similar to Theorem C.2. We note that the expectation is taken over the single uniform permutation τ for PPM-SO and over independent uniform permutations $\tau_1, \tau_2, \ldots, \tau_K$ for PPM-RR.

Step 3: Upper bounding $\mathbb{E}[|\mathbf{r}_k|^2]$ We follow the same procedure as in Step 3 of Theorem C.2 and obtain the following bound by successively applying $\left|\mathbf{J}_{\tilde{\tau}_k(i)}\right| \leq l$ and Young's inequality:

$$|\mathbf{r}_k|^2 \leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 |1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2.$$

We now convert the above inequality into an upper bound for $\mathbb{E}[|\mathbf{r}_k|^2]$ by applying Lemma B.3 to the term $|1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2$. The case of PPM-RR and PPM-SO are discussed separately.

Step 3A: $\mathbb{E}[|\mathbf{r}_k|^2]$ for PPM-RR We write $\mathbb{E}[|\mathbf{r}_k|^2] = \mathbb{E}_{\tau_1,\ldots,\tau_K}[|\mathbf{r}_k|^2]$ to explicitly denote the dependence of the expectation on the uniformly sampled random permutations $\tau_1, \ldots, \tau_K$, and then proceed as follows:

$$\mathbb{E}[|\mathbf{r}_k|^2] \leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \mathbb{E}_{\tau_k}[|1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2].$$

Since τ_k is a uniformly sampled random permutation and $\mathbb{S}_n$ is a finite group with $n!$ elements, it follows that for any $\pi \in \mathbb{S}_n$,

$$\mathbb{P}_{\text{Uniform}}\left[\tau_k = \pi\right] = \mathbb{P}_{\text{Uniform}}\left[\tilde{\tau}_k = \pi\right] = 1/n! \, .$$

Hence, we infer that,

$$\mathbb{E}_{\tau_k}[|1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2] = \mathbb{E}_{\tilde{\tau}_k}[|1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2] = \frac{n-i}{i(n-1)}\sigma_*^2,$$

where the last equality follows from Lemma B.3. Note that σ_*^2 denotes the gradient variance at the minimax point $\mathbf{z}^*$, defined as $\sigma_*^2 = 1/n \sum_{i=1}^{n} |\omega_i(\mathbf{z}^*)|^2$. Substituting the above into the upper bound for $\mathbb{E}[|\mathbf{r}_k|^2]$ yields,

$$\begin{aligned}
\mathbb{E}[|\mathbf{r}_k|^2] &\leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \mathbb{E}_{\tau_k}[|1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2] \\
&\leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 \frac{n-i}{i(n-1)}\sigma_*^2 = l^2 e^{2/5}\sigma_*^2 \sum_{i=1}^{n-1} i(n-i) \\
&\leq l^2 n^3 \sigma_*^2/4.
\end{aligned}$$

Step 3B: $\mathbb{E}[|\mathbf{r}_k|^2]$ for PPM-SO For PPM-SO, $\tau_1 = \ldots = \tau_K = \tau$ where τ is a uniformly sampled random permutation. Consequently, $\tilde{\tau}_1 = \ldots = \tilde{\tau}_K = \tilde{\tau}$ where $\tilde{\tau}$ is the transposed permutation of τ. Hence, we conclude that,

$$|\mathbf{r}_k|^2 \leq l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i^2 |1/i \sum_{j=1}^{i} \omega_{\tilde{\tau}(j)}(\mathbf{z}^*)|^2.$$

35

Proceeding along the same lines as Step 3A, we use the fact that τ is a uniformly sampled random permutation to obtain the following:

$$\mathbb{E}[|\mathbf{r}_k|^2] \leq l^2 e^{2/5}(n-1)\sum_{i=1}^{n-1} i^2 \mathbb{E}_\tau[|1/i\sum_{j=1}^i \omega_{\tilde{\tau}(j)}(\mathbf{z}^*)|^2]$$

$$= l^2 e^{2/5}(n-1)\sum_{i=1}^{n-1} i^2 \mathbb{E}_{\tilde{\tau}}[|1/i\sum_{j=1}^i \omega_{\tilde{\tau}(j)}(\mathbf{z}^*)|^2]$$

$$\leq l^2 e^{2/5}(n-1)\sum_{i=1}^{n-1} i^2 \frac{n-i}{i(n-1)}\sigma_*^2$$

$$\leq l^2 n^3 \sigma_*^2/4.$$

Hence, for both PPM-RR and PPM-SO, $|\mathbf{r}_k|$ is bounded in expectation as $\mathbb{E}[|\mathbf{r}_k|^2] \leq l^2 n^3 \sigma_*^2/4$.

Step 4: Convergence analysis Since $\mathbf{H}_k$ is invertible, we write (D.5) as
$$\mathbf{z}_0^{k+1} - \mathbf{z}^* = \mathbf{H}_k^{-1}[\mathbf{z}_0^k - \mathbf{z}^* - \alpha^2 \mathbf{r}_k].$$

Unrolling for K epochs and setting $\mathbf{z}_0^1 = \mathbf{z}_0$ yields,
$$\mathbf{z}_0^{K+1} - \mathbf{z}^* = \left[\widetilde{\prod}_{k=1}^K \mathbf{H}_k^{-1}\right](\mathbf{z}_0 - \mathbf{z}^*) - \alpha^2 \sum_{k=1}^K \left[\widetilde{\prod}_{j=k}^K \mathbf{H}_j^{-1}\right]\mathbf{r}_k.$$

By the triangle inequality and the bound $|\mathbf{H}_k^{-1}| \leq (1 - n\alpha\mu/2)$, we obtain:
$$\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right| \leq \left[\widetilde{\prod}_{k=1}^K |\mathbf{H}_k^{-1}|\right]|\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^K \left[\widetilde{\prod}_{j=k}^K |\mathbf{H}_j^{-1}|\right]|\mathbf{r}_k|$$

$$\leq (1 - n\alpha\mu/2)^K |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^K (1-n\alpha\mu/2)^{K-k}|\mathbf{r}_k|,$$

and as a result of Young's inequality,
$$\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 \leq 2(1-n\alpha\mu/2)^{2K}|\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^K (1-n\alpha\mu/2)^{2(K-k)}|\mathbf{r}_k|^2.$$

By taking expectations (with respect to the uniform random permutations $\tau_1,\ldots,\tau_K$ for RR and τ for SO respectively) on both sides of the above inequality and substituting the upper bound for $\mathbb{E}[|\mathbf{r}_k|^2]$ for PPM-RR and PPM-SO as derived in Step 3, we obtain

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq 2e^{-n\alpha\mu K}|\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^K (1-n\alpha\mu/2)^{2(K-k)}\mathbb{E}[|\mathbf{r}_k|^2]$$

$$\leq 2e^{-n\alpha\mu K}|\mathbf{z}_0 - \mathbf{z}^*|^2 + 0.5 l^2 n^3 \sigma_*^2 \alpha^4 K \sum_{k=1}^K (1-n\alpha\mu/2)^{2(K-k)}.$$

This yields the following guarantee that holds for PPM-RR and PPM-SO whenever $\alpha \leq \frac{\mu}{5nl^2}$,

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq 2e^{-n\alpha\mu K}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{l^2 \sigma_*^2 \alpha^3 n^2 K}{\mu}. \tag{D.7}$$

We note that the above inequality is identical to the inequality (C.8) obtained in Theorem C.2. Hence, we substitute $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)|\,n^{1/2}K/\mu)/\mu nK\}$ in (D.7) and follow the same steps as in Theorem C.2 to obtain the following last-iterate convergence guarantee for both PPM-RR and PPM-SO, which holds for any $K \geq 1$:

$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] \leq 2e^{-\frac{K}{5\kappa^2}}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 8\kappa^2 \sigma_*^2 \log^3\left(|\nu(\mathbf{z}_0)|\,n^{1/2}K/\mu\right)}{\mu^2 nK^2}. \tag{D.8}$$

Suppressing constant terms and logarithmic factors, we obtain the following:
$$\mathbb{E}[\left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2] = \tilde{O}(e^{-K/5\kappa^2} + 1/nK^2). \tag{D.9}$$

$\square$

36

D.3 Analysis of PPM-AS

Theorem D.3 (Convergence of PPM-AS). *Consider Problem (4) for the μ-strongly monotone operator $\nu(\mathbf{z}) = 1/n \sum_{i=1}^{n} \omega_i(\mathbf{z})$ where each ω_i is l-Lipschitz, but not necessarily monotone. Let $\mathbf{z}^*$ denote the unique root of ν. Then, for any $\alpha \le \frac{\mu}{5nl^2}$ and $K \ge 1$, the iterates of PPM-AS satisfy the following:*

$$\max_{\tau_1,\ldots,\tau_K \in \mathbb{S}_n} \left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 \le 2e^{-n\alpha\mu K} \left|\mathbf{z}_0 - \mathbf{z}^*\right|^2 + \frac{3l^2\sigma_*^2\alpha^3 n^3 K}{\mu}.$$

Setting $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)|K/\mu)/\mu nK\}$ results in the following last-iterate convergence guarantee for PPM-AS, which holds for any $K \ge 1$:

$$\max_{\tau_1,\ldots,\tau_K \in \mathbb{S}_n} \left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 \le 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 24\kappa^2\sigma_*^2\log^3(|\nu(\mathbf{z}_0)|K/\mu)}{\mu^2 K^2}$$

$$= \tilde{O}(e^{-K/5\kappa^2} + 1/K^2).$$

Proof. The iterate-level update rule of PPM-AS can be expressed as

$$\mathbf{z}_i^k = \mathbf{z}_{i-1}^k - \alpha\omega_{\tau_k(i)}(\mathbf{z}_i^k), \tag{D.10}$$

where $k \in [K]$, $i \in [n]$ and τ_k denotes an arbitrary permutation of $[n]$ which is chosen by the adversary at the start of epoch k. For any permutation τ_k chosen by the adversary, we denote the transposed permutation $\tilde{\tau}_k$ as $\tilde{\tau}_k(i) = \tau_k(n+1-i)$. Furthermore $\mathbf{z}_i^k$ is well defined and unique due to the Banach Fixed Point Theorem as $\alpha \le \mu/5nl^2 < 1/l$.

We recall that Steps 1 and 2 of Theorem D.2 are applicable for any permutation $\tau_k \in \mathbb{S}_n$, and hence, can be applied directly to PPM-AS. Consequently, the epoch-level update rule of PPM-AS is the same as that of PPM-RR/SO and is given by:

$$\mathbf{z}_0^k - \mathbf{z}^* = \mathbf{H}_k(\mathbf{z}_0^{k+1} - \mathbf{z}^*) + \alpha^2\mathbf{r}_k, \tag{D.11}$$

where $\mathbf{H}_k$ and $\mathbf{r}_k$ are as defined in Theorem D.2, and are given by:

$$\mathbf{H}_k = \mathbf{I} + \alpha \sum_{j=1}^{n} \left(\widetilde{\prod}_{t=j+1}^{n}(\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)}) \right) \mathbf{M}_{\tilde{\tau}_k(j)},$$

$$\mathbf{r}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^{n}(\mathbf{I} + \alpha\mathbf{J}_{\tilde{\tau}_k(t)}) \right] \mathbf{J}_{\tilde{\tau}_k(i+1)} \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*),$$

where, as before, $\mathbf{M}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_0^{k+1} + (1-t)\mathbf{z}^*)\mathrm{d}t$ and $\mathbf{J}_{\tau_k(i)} = \int_0^1 \nabla\omega_{\tau_k(i)}(t\mathbf{z}_i^k + (1-t)\mathbf{z}_0^{k+1})\mathrm{d}t$. Furthermore, by Step 2 of Theorem D.2, $\mathbf{H}_k$ is invertible and $\left|\mathbf{H}_k^{-1}\right| \le 1 - n\alpha\mu/2$.

We now proceed to control the noise term $\mathbf{r}_k$ by deriving an upper bound for $|\mathbf{r}_k|^2$ that holds uniformly over all permutations $\tau_1,\ldots,\tau_K \in \mathbb{S}_n$. This is done by following the same steps as that of Theorem C.3, which gives us the following bound by successive applications of $\left|\mathbf{J}_{\tilde{\tau}_k(i)}\right| \le l$ and Young's inequality:

37

$$|\mathbf{r}_k|^2 \le l^2 e^{2/5} \left(\sum_{i=1}^{n-1} | \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)| \right)^2$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} | \sum_{j=1}^{i} \omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{i} |\omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2$$

$$\le l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{n} |\omega_{\tilde{\tau}_k(j)}(\mathbf{z}^*)|^2 = l^2 e^{2/5}(n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{n} |\omega_j(\mathbf{z}^*)|^2$$

$$= l^2 e^{2/5}(n-1)(n\sigma_*^2) \sum_{i=1}^{n-1} i \le 3l^2 n^4 \sigma_*^2/4.$$

We note that the above sequence of inequalities hold for any permutation τ_k, and hence, the upper bound $|\mathbf{r}_k|^2 \le 3l^2 n^4 \sigma_*^2/4$ holds for any $\tau_k \in \mathbb{S}_n$, $k \in [K]$.

Since $\mathbf{H}_k$ is invertible, we write (D.11) as

$$\mathbf{z}_0^{k+1} - \mathbf{z}^* = \mathbf{H}_k^{-1}[\mathbf{z}_0^k - \mathbf{z}^* - \alpha^2 \mathbf{r}_k].$$

By unrolling for K epochs and setting $\mathbf{z}_0^1 = \mathbf{z}_0$, we obtain:

$$\mathbf{z}_0^{K+1} - \mathbf{z}^* = \left[\widetilde{\prod}_{k=1}^{K} \mathbf{H}_k^{-1} \right] (\mathbf{z}_0 - \mathbf{z}^*) - \alpha^2 \sum_{k=1}^{K} \left[\widetilde{\prod}_{j=k}^{K} \mathbf{H}_j^{-1} \right] \mathbf{r}_k.$$

By using the triangle inequality and the bound $\left| \mathbf{H}_k^{-1} \right| \le (1 - n\alpha\mu/2)$, we get:

$$\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right| \le \left[\widetilde{\prod}_{k=1}^{K} \left| \mathbf{H}_k^{-1} \right| \right] |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^{K} \left[\widetilde{\prod}_{j=k}^{K} \left| \mathbf{H}_j^{-1} \right| \right] |\mathbf{r}_k|$$

$$\le (1 - n\alpha\mu/2)^K |\mathbf{z}_0 - \mathbf{z}^*| + \alpha^2 \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{K-k} |\mathbf{r}_k|.$$

Applying Young's inequality yields the following:

$$\left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2 \le 2(1 - n\alpha\mu/2)^{2K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} |\mathbf{r}_k|^2$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + 2\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)} |\mathbf{r}_k|^2$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + (1.5l^2 n^4 \sigma_*^2)\alpha^4 K \sum_{k=1}^{K} (1 - n\alpha\mu/2)^{2(K-k)}$$

$$\le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{3l^2 \sigma_*^2 n^3 \alpha^3 K}{\mu}.$$

Since the above inequality holds for any sequence of permutations $\tau_1, \ldots, \tau_K$ chosen by the adversary, we obtain the following uniform guarantee for GDA-AS whenever $\alpha \le \frac{\mu}{5nl^2}$,

$$\max_{\tau_1, \ldots, \tau_K \in \mathbb{S}_n} \left| \mathbf{z}_0^{K+1} - \mathbf{z}^* \right|^2 \le 2e^{-n\alpha\mu K} |\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{3l^2 \sigma_*^2 \alpha^3 n^3 K}{\mu}. \tag{D.12}$$

We note that the above inequality is exactly identical to the inequality (C.16) obtained in Theorem C.3. Hence, we substitute $\alpha = \min\{\mu/5nl^2, 2\log(|\nu(\mathbf{z}_0)| K/\mu)/\mu nK\}$ in (D.12) and follow the

38

same steps as in Theorem C.3 to obtain the following last-iterate convergence guarantee for PPM-AS, which holds for any $K \geq 1$:

$$\max_{\tau_1,\ldots,\tau_K \in \mathbb{S}_n} \left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 \leq 2e^{-K/5\kappa^2}|\mathbf{z}_0 - \mathbf{z}^*|^2 + \frac{2\mu^2 + 24\kappa^2\sigma_*^2 \log^3(|\nu(\mathbf{z}_0)|K/\mu)}{\mu^2 K^2}. \quad \text{(D.13)}$$

Suppressing constant terms and logarithmic factors, we get,

$$\max_{\tau_1,\ldots,\tau_K \in \mathbb{S}_n} \left|\mathbf{z}_0^{K+1} - \mathbf{z}^*\right|^2 = \tilde{O}(e^{-K/5\kappa^2} + 1/K^2). \quad \text{(D.14)}$$

$\square$

E Analysis of AGDA-RR and AGDA-AS

In this section, we present the convergence analysis of AGDA-RR and AGDA-AS for smooth finite-sum objectives that satisfy a two-sided PŁ inequality. We begin by stating some useful properties of two-sided PŁ functions which have been established in prior works and are crucial to our analysis. We then present our convergence analysis of AGDA-RR and AGDA-AS, which follows the same broad structure as our analysis of GDA-RR/AS.

E.1 Properties of Two-sided PŁ Functions

We begin by presenting three important properties of two-sided PŁ functions that are used throughout our analysis of AGDA-RR/AS

Lemma E.1 (Equivalent optimality conditions for two-sided PŁ functions). *Let* $F : \mathbb{R}^{d_{\mathbf{x}}} \times \mathbb{R}^{d_{\mathbf{y}}} \to \mathbb{R}$ *satisfy Assumption 3 and let* $(\mathbf{x}^*, \mathbf{y}^*)$ *be any point in* $\mathbb{R}^{d_{\mathbf{x}}} \times \mathbb{R}^{d_{\mathbf{y}}}$. *Then the following conditions are equivalent:*

- $(\mathbf{x}^*, \mathbf{y}^*)$ *is a global minimax point, i.e.,* $\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^{d_{\mathbf{x}}}} \Phi(\mathbf{x})$ *where* Φ *is the best response function defined by* $\Phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}} F(\mathbf{x}, \mathbf{y})$, *and* $\mathbf{y}^* \in \arg\max_{\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}} F(\mathbf{x}^*, \mathbf{y})$.

- $(\mathbf{x}^*, \mathbf{y}^*)$ *is a saddle point, i.e.,*

$$F(\mathbf{x}^*, \mathbf{y}) \leq F(\mathbf{x}^*, \mathbf{y}^*) \leq F(\mathbf{x}, \mathbf{y}^*), \; \forall \mathbf{x} \in \mathbb{R}^{d_{\mathbf{x}}}, \mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}.$$

- $(\mathbf{x}^*, \mathbf{y}^*)$ *is a stationary point, i.e.,*

$$\nabla_{\mathbf{x}} F(\mathbf{x}^*, \mathbf{y}^*) = \nabla_{\mathbf{y}} F(\mathbf{x}^*, \mathbf{y}^*) = 0.$$

Proof. See Lemma 2.1 of Yang et al. [54]. $\square$

Lemma E.2 (Properties of the best response). *Let* $F : \mathbb{R}^{d_{\mathbf{x}}} \times \mathbb{R}^{d_{\mathbf{y}}} \to \mathbb{R}$ *be an l-smooth function satisfying Assumption 3 (with constants* μ_1 *and* μ_2) *and let* $\Phi : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}$ *be the best response function defined as* $\Phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}} F(\mathbf{x}, \mathbf{y})$. *Then* Φ *satisfies the following properties:*

- Φ *is differentiable and L-smooth where* $L = l + l^2/\mu_2$.

- $\nabla\Phi(\mathbf{x}) = \nabla_{\mathbf{x}} F(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ *where* $\mathbf{y}^*(\mathbf{x})$ *is any arbitrary point in* $\arg\max_{\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}} F(\mathbf{x}, \mathbf{y})$.

- Φ *satisfies the (one-sided) PŁ inequality with constant* μ_1, *i.e.,*

$$|\nabla\Phi(\mathbf{x})|^2 \geq 2\mu_1[\Phi(\mathbf{x}) - \Phi^*],$$

where $\Phi^* = \min_{\mathbf{x} \in \mathbb{R}^{d_{\mathbf{x}}}} \Phi(\mathbf{x})$.

Proof. See Lemma A.2 and Lemma A.3 of Yang et al. [54]. $\square$

Lemma E.3 (Quadratic growth properties of PŁ functions). *Let* $f : \mathbb{R}^{d_{\mathbf{x}}} \to \mathbb{R}$ *be a PŁ function, i.e., there exists a positive constant* μ *such that* $|\nabla f(\mathbf{x})|^2 \geq 2\mu[f(\mathbf{x}) - f^*]$, *where* $f^* = \min_{\mathbf{x}} f(\mathbf{x})$. *Then,* f *satisfies the quadratic growth property with constant* μ, *i.e., for any* $\mathbf{x} \in \mathbb{R}^{d_{\mathbf{x}}}$, $f(\mathbf{x}) - f^* \geq \frac{\mu}{2}|\mathbf{x}^* - \mathbf{x}|^2$ *where* $\mathbf{x}^*$ *is the projection of* $\mathbf{x}$ *on the set* $\arg\min_{\mathbf{x}} f(\mathbf{x})$.

39

Proof. See Lemma A.1 of Yang et al. [54]. □

Finally, we motivate the necessity of the bounded gradient variance assumption by presenting the construction of a two-sided PŁ function whose set of global minimax points is unbounded.

Lemma E.4 (2PŁ function with unbounded set of global minimax points). *Consider the function* $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ *given by* $f(\mathbf{x}, \mathbf{y}) = (\mathbf{x}_1 + \mathbf{x}_2)^2/2 - (\mathbf{y}_1 + \mathbf{y}_2)^2/2$. *Then,* f *is a two-sided PŁ function with constants* $\mu_1 = \mu_2 = 2$. *Furthermore, the set of global minimax points of* F *is an unbounded proper subset of* $\mathbb{R}^4$.

Proof. We define the functions $\Phi : \mathbb{R}^2 \to \mathbb{R}$ and $\Psi : \mathbb{R}^2 \to \mathbb{R}$ as follows:

$$\Phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^2} f(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x}_1 + \mathbf{x}_2)^2}{2},$$

$$\Psi(\mathbf{y}) = \min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}, \mathbf{y}) = -\frac{(\mathbf{y}_1 + \mathbf{y}_2)^2}{2}.$$

Furthermore, we note that:

$$\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = [\mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_1 + \mathbf{x}_2],$$
$$\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = -[\mathbf{y}_1 + \mathbf{y}_2, \mathbf{y}_1 + \mathbf{y}_2].$$

From the above inequalities, it follows that:

$$|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})|^2 = 4[f(\mathbf{x}, \mathbf{y}) - \Psi(\mathbf{y})] = 4[f(\mathbf{x}, \mathbf{y}) - \min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}, \mathbf{y})],$$

$$|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})|^2 = 4[\Phi(\mathbf{x}) - f(\mathbf{x}, \mathbf{y})] = 4[\max_{\mathbf{y} \in \mathbb{R}^2} f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y})].$$

Thus, f satisfies the two-sided PŁ inequality with constants $\mu_1 = \mu_2 = 2$. Consequently, Lemma E.1 implies that the set of minimax points are exactly the ones satisfying $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$, i.e., the set of global minimax points is $\{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^4 \mid \mathbf{x}_1 + \mathbf{x}_2 = 0, \mathbf{y}_1 + \mathbf{y}_2 = 0\}$, which is an unbounded proper subset of $\mathbb{R}^4$. □

E.2 Analysis of AGDA-RR

Theorem E.1 (Convergence of AGDA-RR). *Let Assumptions 1, 3, and 4 be satisfied and let* $\eta = {}^{73l^2}\!/2\mu_2^2$. *Then, for any* $\alpha \leq {}^1\!/5\eta nl$, $\beta = \eta\alpha$, *and* $K \geq 1$, *the iterates of AGDA-RR satisfy the following:*

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + 0.42n\alpha^2 l^2 \sigma^2 (3 + \eta^3)/\mu_1.$$

Setting $\alpha = \min\{1/5\eta nl, 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/\mu_1 nK\}$ *results in the following expected last iterate convergence guarantee, which holds for any* $K \geq 1$:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq e^{-K/365\kappa^3} V_\lambda(\mathbf{z}_0) + \frac{\mu_1 + c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{\mu_1 nK^2}$$

$$= \tilde{O}(e^{-K/365\kappa^3} + 1/nK^2),$$

where $\kappa = \max\{{}^l\!/\mu_1, {}^l\!/\mu_2\}$ *and* $c > 0$ *is a numerical constant that is independent of* l, μ_1, μ_2 *and* σ^2.

Proof. Before presenting the convergence analysis of AGDA-RR, we introduce some additional notation. For any $k \in [K]$ and any random variable Y, we use $\mathbb{E}_k[Y]$ to denote the expectation of Y conditioned on all previous epoch iterates $(\mathbf{x}_0^j, \mathbf{y}_0^j)_{j=1}^k$, i.e., $\mathbb{E}_k[Y] = \mathbb{E}[Y \mid \mathbf{x}_0^1, \mathbf{y}_0^1, \ldots, \mathbf{x}_0^k, \mathbf{y}_0^k]$. Similarly, for any two random variables Y, Z, we use $\mathbb{E}_k[Y \mid Z]$ to denote the expectation of Y conditioned on all previous epoch iterates $(\mathbf{x}_0^j, \mathbf{y}_0^j)_{j=1}^k$ and Z, i.e. $\mathbb{E}_k[Y \mid Z] = \mathbb{E}[Y \mid \mathbf{x}_0^1, \mathbf{y}_0^1, \ldots, \mathbf{x}_0^k, \mathbf{y}_0^k, Z]$.

Our analysis follows a structure similar to that of GDA-RR/SO and consists of three major steps: 1. obtaining an epoch level update rule for AGDA-RR that resembles the update of deterministic AGDA with noise, 2. bounding the noise in the update rule using Lemma B.3, 3. presenting a convergence analysis with respect to the Lyapunov function V_λ.

40

Step 1: Epoch level update rule We begin with the observation that the x-iterates of AGDA-RR are the same as that of gradient descent without replacement on $F(\mathbf{x}, \mathbf{y}_0^k)$ and thus, can be treated as a noisy version of full-batch gradient descent. To this end, we linearize the stochastic gradients $\nabla_{\mathbf{x}} f_{\tau_k(i)}(\mathbf{x}_{i-1}^k, \mathbf{y}_0^k)$ about $(\mathbf{x}_0^k, \mathbf{y}_0^k)$ and obtain the following decomposition:

$$\nabla_{\mathbf{x}} f_{\tau_k(i)}(\mathbf{x}_{i-1}^k, \mathbf{y}_0^k) = \nabla_{\mathbf{x}} f_{\tau_k(i)}(\mathbf{x}_0^k, \mathbf{y}_0^k) + \left[\int_0^1 \nabla_{\mathbf{xx}}^2 f_{\tau_k(i)}(\mathbf{x}_0^k + t(\mathbf{x}_{i-1}^k - \mathbf{x}_0^k), \mathbf{y}_0^k)dt\right](\mathbf{x}_{i-1}^k - \mathbf{x}_0^k)$$

$$= \nabla_{\mathbf{x}} f_{\tau_k(i)}(\mathbf{x}_0^k, \mathbf{y}_0^k) + \mathbf{H}_{\tau_k(i)}(\mathbf{x}_{i-1}^k - \mathbf{x}_0^k),$$

where $\mathbf{H}_{\tau_k(i)} = \int_0^1 \nabla_{\mathbf{xx}}^2 f_{\tau_k(i)}(\mathbf{x}_0^k + t(\mathbf{x}_{i-1}^k - \mathbf{x}_0^k), \mathbf{y}_0^k)dt$. As before, by Rademacher's Theorem, the l-smoothness of $f_{\tau_k(i)}$ guarantees that $\mathbf{H}_{\tau_k(i)}$ is well defined and satisfies $\left|\mathbf{H}_{\tau_k(i)}\right| \leq l$.

When substituting the above linearization into the x-iterate updates, we observe the following:

$$\mathbf{x}_1^k = \mathbf{x}_0^k - \alpha \nabla_{\mathbf{x}} f_{\tau_k(1)}(\mathbf{x}_0^k, \mathbf{y}_0^k),$$

$$\mathbf{x}_2^k = \mathbf{x}_1^k - \alpha \nabla_{\mathbf{x}} f_{\tau_k(2)}(\mathbf{x}_1^k, \mathbf{y}_0^k) = \mathbf{x}_1^k - \alpha \nabla_{\mathbf{x}} f_{\tau_k(2)}(\mathbf{x}_0^k, \mathbf{y}_0^k) - \alpha \mathbf{H}_{\tau_k(2)}(\mathbf{x}_1^k - \mathbf{x}_0^k)$$

$$= \mathbf{x}_0^k - \alpha \nabla_{\mathbf{x}} f_{\tau_k(2)}(\mathbf{x}_0^k, \mathbf{y}_0^k) - \alpha \nabla_{\mathbf{x}} f_{\tau_k(1)}(\mathbf{x}_0^k, \mathbf{y}_0^k) + \alpha^2 \mathbf{H}_{\tau_k(2)} \nabla_{\mathbf{x}} f_{\tau_k(1)}(\mathbf{x}_0^k, \mathbf{y}_0^k)$$

$$= \mathbf{x}_0^k - \alpha(\mathbf{I} - \alpha \mathbf{H}_{\tau_k(2)})\nabla_{\mathbf{x}} f_{\tau_k(1)}(\mathbf{x}_0^k, \mathbf{y}_0^k) - \alpha \nabla_{\mathbf{x}} f_{\tau_k(2)}(\mathbf{x}_0^k, \mathbf{y}_0^k).$$

Repeating the above steps for subsequent iterates $\mathbf{x}_3^k, \ldots, \mathbf{x}_n^k$, we obtain the following epoch-level update rule for $\mathbf{x}_0^{k+1} = \mathbf{x}_n^k$:

$$\mathbf{x}_0^{k+1} = \mathbf{x}_0^k - \alpha \sum_{j=1}^n \left(\widetilde{\prod}_{t=j+1}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)})\right) \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k). \tag{E.1}$$

We further simplify (E.1) using the summation by parts identity. To this end, we define c_j and e_j as follows:

$$c_j = \widetilde{\prod}_{t=j+1}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)}),$$

$$e_j = \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k).$$

We note that $c_n = \mathbf{I}$, since $\widetilde{\prod}_{t=n+1}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)})$ denotes the empty reverse product. Hence, we conclude,

$$c_n \sum_{j=1}^n e_j = \sum_{j=1}^n \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k) = n \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k),$$

$$\sum_{i=1}^{n-1} (c_{i+1} - c_i) \sum_{j=1}^i e_j = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)}) - \widetilde{\prod}_{t=i+1}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)})\right] \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)$$

$$= \alpha \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)})\right] \mathbf{H}_{\tau_k(i+1)} \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k).$$

We define $\mathbf{r}_k$ as,

$$\mathbf{r}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)})\right] \mathbf{H}_{\tau_k(i+1)} \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k). \tag{E.2}$$

From the summation by parts identity $\sum_{i=1}^n c_i e_i = c_n \sum_{j=1}^n e_j - \sum_{i=1}^{n-1} (c_{i+1} - c_i) \sum_{j=1}^i e_j$, we obtain the following:

$$\sum_{j=1}^n \left[\widetilde{\prod}_{t=j+1}^n [\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)}]\right] \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k) = n \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \alpha \mathbf{r}_k.$$

By substituting the above equation into (E.1), we obtain the following update rule for the epoch iterates $\mathbf{x}_0^k$:

$$\mathbf{x}_0^{k+1} = \mathbf{x}_0^k - n\alpha \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) + \alpha^2 \mathbf{r}_k. \tag{E.3}$$

41

The epoch-level update rule for the $\mathbf{y}$-iterate is derived in a similar fashion. We begin by linearizing $\nabla_{\mathbf{y}} f_{\pi_k(i)}(\mathbf{x}_0^{k+1}, \mathbf{y}_{i-1}^k)$ as:

$$\nabla_{\mathbf{y}} f_{\pi_k(i)}(\mathbf{x}_0^{k+1}, \mathbf{y}_{i-1}^k) = \nabla_{\mathbf{y}} f(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) + \mathbf{J}_{\pi_k(i)}(\mathbf{y}_{i-1}^k - \mathbf{y}_0^k),$$

where $\mathbf{J}_{\pi_k(i)} = \int_0^1 \nabla_{\mathbf{yy}}^2 f_{\pi_k(i)}(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k + t(\mathbf{y}_{i-1}^k - \mathbf{y}_0^k)) \mathrm{d}t$ satisfies $\left| \mathbf{J}_{\pi_k(i)} \right| \leq l$. As before, by substituting into the $\mathbf{y}$ iterate updates and unrolling, we obtain,

$$\mathbf{y}_0^{k+1} = \mathbf{y}_0^k + \beta \sum_{j=1}^n \left[\widetilde{\prod}_{t=j+1}^n [\mathbf{I} + \beta \mathbf{J}_{\pi_k(t)}] \right] \nabla_{\mathbf{y}} f_{\pi_k(j)}(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k). \tag{E.4}$$

Once again, applying the summation by parts technique gives us the following update rule for the epoch iterates $\mathbf{y}_0^k$:

$$\mathbf{y}_0^{k+1} = \mathbf{y}_0^k + n\beta \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) + \beta^2 \mathbf{s}_k, \tag{E.5}$$

where $\mathbf{s}_k$ is given by,

$$\mathbf{s}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} + \beta \mathbf{J}_{\pi_k(t)}) \right] \mathbf{J}_{\pi_k(i+1)} \sum_{j=1}^i \nabla_{\mathbf{y}} f_{\pi_k(j)}(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k). \tag{E.6}$$

Step 2: Controlling the noise terms By applying the triangle inequality on (E.2) and substituting $|\mathbf{H}_{\tau_k}(i)| \leq l$ and $\alpha \leq 1/5nl$ wherever required, we bound $|\mathbf{r}_k|$ as follows:

$$|\mathbf{r}_k| \leq \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^n |\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)}| \right] |\mathbf{H}_{\tau_k(i+1)}| \, | \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|$$

$$\leq l(1+\alpha l)^n \sum_{i=1}^{n-1} | \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)| \leq le^{1/5} \sum_{i=1}^{n-1} | \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|$$

$$\leq le^{1/5} \sum_{i=1}^{n-1} i |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) + 1/i \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|$$

$$\leq le^{1/5} \left[n^2 \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| /2 + \sum_{i=1}^{n-1} i |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - 1/i \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)| \right].$$

By defining $d_{k,i} = |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - 1/i \sum_{j=1}^i \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|$ and substituting into the above inequality, we obtain,

$$|\mathbf{r}_k| \leq le^{1/5} \left[n^2 \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| /2 + \sum_{i=1}^{n-1} i d_{k,i} \right]. \tag{E.7}$$

By applying Young's inequality to (E.7), we conclude:

$$|\mathbf{r}_k|^2 \leq 2l^2 e^{2/5} \left[n^4 \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 /4 + (n-1) \sum_{i=1}^{n-1} i^2 d_{k,i}^2 \right]. \tag{E.8}$$

By Lemma B.3 and the bounded gradient variance assumption, we bound $\mathbb{E}[d_{k,i}^2]$ as follows:

$$\mathbb{E}[d_{k,i}^2] = \frac{n-i}{i(n-1)} \left[1/n \sum_{i=1}^n |\nabla_{\mathbf{x}} f_i(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)|^2 \right] \leq \frac{n-i}{i(n-1)} \sigma^2, \tag{E.9}$$

where the expectation is taken over the uniformly sampled random permutation τ_k. Furthermore, applying Jensen's inequality to (E.9) gives us the following bound:

$$\mathbb{E}[d_{k,i}] \leq \sqrt{\mathbb{E}[d_{k,i}^2]} \leq \sigma \sqrt{\frac{n-i}{i(n-1)}}. \tag{E.10}$$

42

From (E.10), it follows that,

$$\sum_{i=1}^{n-1} i\mathbb{E}[d_{k,i}] \leq \sigma \sum_{i=1}^{n-1} \sqrt{i}\sqrt{\frac{n-i}{n-1}} \leq \sigma \sum_{i=1}^{n-1} \sqrt{i} \leq \frac{2\sigma n^{3/2}}{3}.$$

By taking expectations of (E.7) and substituting the above, we get,

$$\mathbb{E}[|\mathbf{r}_k|] \leq \frac{2n^2 l}{3} \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right| + \sigma l n^{3/2}. \tag{E.11}$$

Similarly, from (E.9), it follows that,

$$(n-1)\sum_{i=1}^{n-1} i^2 \mathbb{E}[d_{k,i}^2] \leq \sigma^2 \sum_{i=1}^{n-1} i(n-i) \leq \frac{\sigma^2 n^3}{6}.$$

By taking expectations of (E.8) and substituting the above, we obtain,

$$\mathbb{E}[|\mathbf{r}_k|^2] \leq \frac{3n^4 l^2}{4} \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + \frac{\sigma^2 l^2 n^3}{2}. \tag{E.12}$$

We bound $\mathbb{E}[|\mathbf{s}_k|]$ and $\mathbb{E}[|\mathbf{s}_k|^2]$ by following a similar strategy and obtain:

$$\mathbb{E}[|\mathbf{s}_k|] \leq \frac{2n^2 l}{3} \left|\nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right| + \sigma l n^{3/2}, \tag{E.13}$$

$$\mathbb{E}[|\mathbf{s}_k|^2] \leq \frac{3n^4 l^2}{4} \left|\nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right|^2 + \frac{\sigma^2 l^2 n^3}{2}. \tag{E.14}$$

Step 3: Obtaining a convergence guarantee We begin by defining a_k and b_k as:

$$a_k = \Phi(\mathbf{x}_0^k) - \Phi^*,$$
$$b_k = \Phi(\mathbf{x}_0^k) - F(\mathbf{x}_0^k, \mathbf{y}_0^k).$$

By definition of V_λ, we conclude that $V_\lambda(\mathbf{z}_0^k) = a_k + \lambda b_k$. Furthermore, we note that by Lemma E.2, Φ is L-smooth where $L = l + l^2/\mu_2$. Since $\alpha \leq 1/5\eta nl$, it follows that $n\alpha L \leq 1$. Using these properties, we obtain the following:

$$\Phi(\mathbf{x}_0^{k+1}) - \Phi(\mathbf{x}_0^k) \leq \left\langle \nabla\Phi(\mathbf{x}_0^k), \mathbf{x}_0^{k+1} - \mathbf{x}_0^k \right\rangle + \frac{L}{2}\left|\mathbf{x}_0^{k+1} - \mathbf{x}_0^k\right|^2$$

$$\leq \left\langle \nabla\Phi(\mathbf{x}_0^k), -n\alpha\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) + \alpha^2\mathbf{r}_k \right\rangle + \frac{L}{2}\left|-n\alpha\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) + \alpha^2\mathbf{r}_k\right|^2$$

$$\leq -n\alpha\left\langle \nabla\Phi(\mathbf{x}_0^k), \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \frac{\alpha}{n}\mathbf{r}_k \right\rangle + \frac{L}{2}n^2\alpha^2\left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \frac{\alpha}{n}\mathbf{r}_k\right|^2$$

$$\leq \frac{n\alpha}{2}\left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k) - \frac{\alpha}{n}\mathbf{r}_k\right|^2 - \frac{n\alpha}{2}\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$

$$\leq n\alpha\left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 - \frac{n\alpha}{2}\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2 + \frac{\alpha^3}{n}\left|\mathbf{r}_k\right|^2.$$

By taking expectations conditioned on the epoch iterates $(\mathbf{x}_0^j, \mathbf{y}_0^j)_{j=1}^k$ and substituting the bound (E.12), we obtain the following:

$$\mathbb{E}_k[\Phi(\mathbf{x}_0^{k+1})] - \Phi(\mathbf{x}_0^k) \leq n\alpha\left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 - \frac{n\alpha}{2}\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$

$$+ \frac{3n^3\alpha^3 l^2}{4}\left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + \frac{n^2\alpha^3 l^2\sigma^2}{2}. \tag{E.15}$$

43

We proceed by applying the l-smoothness of $-F(\mathbf{x}, \mathbf{y})$ in $\mathbf{x}$, to obtain the following:

$$F(\mathbf{x}_0^k, \mathbf{y}_0^k) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \leq -\langle \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k), \mathbf{x}_0^{k+1} - \mathbf{x}_0^k \rangle + \frac{l}{2} \left| \mathbf{x}_0^{k+1} - \mathbf{x}_0^k \right|^2$$

$$\leq \langle \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k), n\alpha \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \alpha^2 \mathbf{r}_k \rangle$$

$$+ \frac{l}{2} \left| n\alpha \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \alpha^2 \mathbf{r}_k \right|^2$$

$$\leq (n\alpha + \frac{l}{2} n^2 \alpha^2) \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{l}{2} \left| \alpha^2 \mathbf{r}_k \right|^2$$

$$- (\alpha^2 + l n \alpha^3) \langle \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k), \mathbf{r}_k \rangle$$

$$\leq (n\alpha + \frac{l}{2} n^2 \alpha^2) \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{l\alpha^4}{2} \left| \mathbf{r}_k \right|^2$$

$$+ \frac{6\alpha^2}{5} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| \left| \mathbf{r}_k \right|. \tag{E.16}$$

From (E.11) and (E.12), we infer that:

$$\frac{l\alpha^4}{2} \mathbb{E}[|\mathbf{r}_k|^2] \leq \frac{3n^4 \alpha^4 l^3}{8} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{\sigma^2 l^3 \alpha^4 n^3}{4}$$

$$\leq \frac{3n^4 \alpha^4 l^3}{8} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{\sigma^2 l^2 \alpha^3 n^2}{20},$$

$$\frac{6\alpha^2}{5} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| \mathbb{E}[|\mathbf{r}_k|] \leq \frac{6\alpha^2 n^{\frac{3}{2}} \sigma l}{5} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| + \frac{4n^2 \alpha^2 l}{5} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2$$

$$= (\frac{6\alpha^{\frac{3}{2}} n l \sigma}{5})(\alpha^{\frac{1}{2}} n^{\frac{1}{2}} \left| \nabla_{\mathbf{x}} F(x_0^k, y_0^k) \right|) + \frac{4n^2 \alpha^2 l}{5} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2$$

$$\leq (\frac{n\alpha}{2} + \frac{4n^2 \alpha^2 l}{5}) \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{18\alpha^3 n^2 l^2 \sigma^2}{25}.$$

Taking expectations on both sides of (E.16) by conditioning on the epoch iterates $(\mathbf{x}_0^j, \mathbf{y}_0^j)_{j=1}^k$ and substituting the above inequalities, we obtain:

$$F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \mathbb{E}_k[F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)] \leq (\frac{3n\alpha}{2} + \frac{13}{10} l n^2 \alpha^2 + \frac{3n^4 \alpha^4 l^3}{8}) \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{4\alpha^3 n^2 l^2 \sigma^2}{5}. \tag{E.17}$$

We now apply the l-smoothness of $-F(\mathbf{x}, \mathbf{y})$ in $\mathbf{y}$ to obtain the following:

$$F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1}) \leq -\langle \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k), \mathbf{y}_0^{k+1} - \mathbf{y}_0^k \rangle + \frac{l}{2} \left| \mathbf{y}_0^{k+1} - \mathbf{y}_0^k \right|^2$$

$$\leq -\langle \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k), n\beta \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) + \beta^2 \mathbf{s}_k \rangle$$

$$+ \frac{l}{2} \left| n\beta \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) + \beta^2 \mathbf{s}_k \right|^2$$

$$\leq (-n\beta + l n^2 \beta^2) \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|^2$$

$$+ \beta^2 \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right| \left| \mathbf{s}_k \right| + l \beta^4 \left| \mathbf{s}_k \right|^2. \tag{E.18}$$

From (E.13) and (E.14), we conclude,

$$l\beta^4 \mathbb{E}[|\mathbf{s}_k|^2] \leq \frac{3n^4 l^3 \beta^4}{4} \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|^2 + \frac{\sigma^2 l^3 n^3 \beta^4}{2}$$

$$\leq \frac{3n^4 l^3 \beta^4}{4} \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|^2 + \frac{\sigma^2 l^2 n^2 \beta^3}{10},$$

$$\beta^2 \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right| \mathbb{E}[|\mathbf{s}_k|] \leq \sigma l n^{3/2} \beta^2 \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right| + \frac{2n^2 \beta^2 l}{3} \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|^2$$

$$\leq (\frac{n^{\frac{1}{2}} \beta^{\frac{1}{2}}}{2} \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|)(2n\beta^{\frac{3}{2}} l\sigma) + \frac{2n^2 \beta^2 l}{3} \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|^2$$

$$\leq (\frac{n\beta}{8} + \frac{2n^2 \beta^2 l}{3}) \left| \nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \right|^2 + 2n^2 \beta^3 l^2 \sigma^2.$$

44

We proceed by taking expectations on both sides of (E.18) by conditioning on the epoch iterates $(\mathbf{x}_0^j, \mathbf{y}_0^j)_{j=1}^k$ and $\mathbf{x}_0^{k+1}$ and substituting the above bounds. Consequently, we obtain the following inequality:

$$F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) - \mathbb{E}_k[F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1})|\mathbf{x}_0^{k+1}] \leq (-\frac{7n\beta}{8} + \frac{5n^2\beta^2 l}{3} + \frac{3n^4\beta^4 l^3}{4}) \left|\nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right|^2$$

$$+ \frac{21n^2\beta^3 l^2\sigma^2}{10}$$

$$\leq -\frac{n\beta}{2} \left|\nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right|^2 + \frac{21n^2\beta^3 l^2\sigma^2}{10}.$$

The two-sided PŁ condition on F implies that,

$$-\frac{n\beta}{2} \left|\nabla_{\mathbf{y}} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right|^2 \leq -n\beta\mu_2[\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)],$$

which gives us the following:

$$\mathbb{E}_k[\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1})|\mathbf{x}_0^{k+1}] \leq (1 - n\beta\mu_2)[\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)] + \frac{21n^2\beta^3 l^2\sigma^2}{10}.$$

Taking expectations of the above with respect to $\mathbf{x}_0^{k+1}$ and applying the law of iterated expectation, we obtain:

$$\mathbb{E}_k[\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1})] \leq (1 - n\beta\mu_2)\mathbb{E}_k[\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)] + \frac{21n^2\beta^3 l^2\sigma^2}{10}. \quad \text{(E.19)}$$

From (E.15), we obtain the following:

$$\mathbb{E}_k[a_{k+1}] \leq a_k + n\alpha \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 - \frac{n\alpha}{2} \left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$

$$+ \frac{3n\alpha}{100} \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + \frac{n^2\alpha^3 l^2\sigma^2}{2}. \quad \text{(E.20)}$$

Furthermore, we note that,

$$\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) = b_k + F(\mathbf{x}_0^k, \mathbf{y}_0^k) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) + \Phi(\mathbf{x}_0^{k+1}) - \Phi(\mathbf{x}_0^k).$$

By substituting the above into (E.19), we obtain the following:

$$\mathbb{E}_k[b_{k+1}] \leq (1 - n\beta\mu_2)b_k + (1 - n\beta\mu_2)\mathbb{E}_k[F(\mathbf{x}_0^k, \mathbf{y}_0^k) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)]$$

$$+ (1 - n\beta\mu_2)\mathbb{E}_k[\Phi(\mathbf{x}_0^{k+1}) - \Phi(\mathbf{x}_0^k)] + \frac{21n^2\beta^3 l^2\sigma^2}{10}.$$

By substituting (E.17) and (E.15) into the above, we obtain:

$$\mathbb{E}_k[b_{k+1}] \leq (1 - n\beta\mu_2)b_k + (1 - n\beta\mu_2)n\alpha \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2$$

$$- (1 - n\beta\mu_2)\frac{n\alpha}{2} \left|\nabla\Phi(\mathbf{x}_0^k)\right|^2 + (1 - n\beta\mu_2)\frac{9n\alpha}{5} \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2$$

$$+ (1 - n\beta\mu_2)\frac{13n^2\alpha^3 l^2\sigma^2}{10} + \frac{21n^2\beta^3 l^2\sigma^2}{10}. \quad \text{(E.21)}$$

From (E.20) and (E.21), we get the following by substituting $V_\lambda(\mathbf{z}_0^{k+1}) = a_{k+1} + \lambda b_{k+1}$,

$$\mathbb{E}_k[V_\lambda(\mathbf{z}_0^{k+1})] \leq a_k + (1 - n\beta\mu_2)\lambda b_k + [n\alpha + \lambda n\alpha(1 - n\beta\mu_2)] \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 + C_3$$

$$- [\frac{n\alpha}{2} + \frac{\lambda n\alpha}{2}(1 - n\beta\mu_2)] \left|\nabla\Phi(\mathbf{x}_0^k)\right|^2 + [\frac{3n\alpha}{100} + \lambda(1 - n\beta\mu_2)\frac{9n\alpha}{5}] \left|\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2, \quad \text{(E.22)}$$

where C_3 is defined as follows:

$$C_3 = [\frac{1}{2} + \frac{13\lambda}{10}(1 - n\beta\mu_2)]n^2\alpha^3 l^2\sigma^2 + \frac{21\lambda n^2\beta^3 l^2\sigma^2}{10}.$$

45

By Young's inequality, the following holds for any $\varepsilon \in (0, 1]$:

$$\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 \leq (1 + \varepsilon)\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2 + (1 + 1/\varepsilon)\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2.$$

Substituting this into (E.22), we obtain the following:

$$\mathbb{E}_k[V_\lambda(\mathbf{z}_0^{k+1})] \leq a_k + (1 - n\beta\mu_2)\lambda b_k + C_3$$
$$- [\frac{n\alpha}{2} - (1 + \varepsilon)\frac{3n\alpha}{100} - \lambda(1 - n\beta\mu_2)[(1 + \varepsilon)\frac{9n\alpha}{5} - \frac{n\alpha}{2}]]\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$
$$+ [n\alpha + (1 + 1/\varepsilon)\frac{3n\alpha}{100} + \lambda(1 - n\beta\mu_2)[n\alpha + (1 + 1/\varepsilon)\frac{9n\alpha}{5}]]\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2.$$
$$\text{(E.23)}$$

From Lemma E.2, we know that $\Phi(\mathbf{x})$ satisfies the PŁ inequality with constant μ_1. Thus, we conclude that

$$\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2 \geq 2\mu_1(\Phi(\mathbf{x}_0^k) - \Phi^*) = 2\mu_1 a_k. \qquad \text{(E.24)}$$

Let $\mathbf{y}^*(\mathbf{x}_0^k)$ denote the projection of $\mathbf{y}_0^k$ on the set $\arg\max_{\mathbf{y}} F(\mathbf{x}_0^k, \mathbf{y})$. From Lemma E.2, we conclude that $\nabla\Phi(\mathbf{x}_0^k) = \nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}^*(\mathbf{x}_0^k))$. Thus,

$$\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 \leq \left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}^*(\mathbf{x}_0^k))\right|^2$$
$$\leq l^2\left|\mathbf{y}_0^k - \mathbf{y}^*(\mathbf{x}_0^k)\right|^2 \qquad \text{(E.25)}$$

Furthermore, we note that Assumption 3 implies that $-F(\mathbf{x}_0^k, \mathbf{y})$ satisfies the PŁ inequality with constant μ_2. Thus, by Lemma E.3, we conclude that,

$$b_k = \Phi(\mathbf{x}_0^k) - F(\mathbf{x}_0^k, \mathbf{y}_0^k) \geq \frac{\mu_2}{2}\left|\mathbf{y}_0^k - \mathbf{y}^*(\mathbf{x}_0^k)\right|^2.$$

By substituting the above inequality in (E.25), we conclude that,

$$\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 \leq \frac{2l^2 b_k}{\mu_2}.$$

We proceed by substituting the above bound in (E.23) to obtain the following upper bound on $\mathbb{E}_k[V_\lambda(\mathbf{z}_0^{k+1})]$:

$$\mathbb{E}_k[V_\lambda(\mathbf{z}_0^{k+1})] \leq \{1 - [\mu_1 n\alpha - (1 + \varepsilon)\frac{3\mu_1 n\alpha}{50} - \lambda(1 - n\beta\mu_2)[(1 + \varepsilon)\frac{18\mu_1 n\alpha}{5} - \mu_1 n\alpha]]\}a_k + C_3$$
$$+ [1 - n\beta\mu_2 + \frac{2l^2 n\alpha}{\mu_2\lambda} + (1 + 1/\varepsilon)\frac{3l^2 n\alpha}{50\mu_2\lambda} + (1 - n\beta\mu_2)[\frac{2l^2 n\alpha}{\mu_2} + (1 + 1/\varepsilon)\frac{18l^2 n\alpha}{5\mu_2}]]\lambda b_k.$$

Hence, we obtain the following update equation for $V_\lambda(\mathbf{z}_0^{k+1})$:

$$\mathbb{E}_k[V_\lambda(\mathbf{z}_0^{k+1})] \leq C_1 a_k + C_2\lambda b_k + C_3, \qquad \text{(E.26)}$$

where C_1, C_2 and C_3 are given by,

$$C_1 = 1 - [\mu_1 n\alpha - (1 + \varepsilon)\frac{3\mu_1 n\alpha}{50} - \lambda(1 - n\beta\mu_2)[(1 + \varepsilon)\frac{18\mu_1 n\alpha}{5} - \mu_1 n\alpha]]$$
$$= 1 - \mu_1 n\alpha[1 - \frac{3(1 + \varepsilon)}{50} - \lambda(1 - n\beta\mu_2)[\frac{18(1 + \varepsilon)}{5} - 1]],$$
$$C_2 = 1 - n\beta\mu_2 + \frac{2l^2 n\alpha}{\mu_2\lambda} + (1 + 1/\varepsilon)\frac{3l^2 n\alpha}{50\mu_2\lambda} + (1 - n\beta\mu_2)[\frac{2l^2 n\alpha}{\mu_2} + (1 + 1/\varepsilon)\frac{18l^2 n\alpha}{5\mu_2}]$$
$$= 1 - \frac{2l^2 n\alpha}{\mu_2}[\frac{\beta\mu_2^2}{2\alpha l^2} - \frac{1}{\lambda} - (1 + 1/\varepsilon)\frac{3}{100\lambda} - (1 - n\beta\mu_2)[1 + (1 + 1/\varepsilon)\frac{9}{5}]],$$
$$C_3 = [\frac{1}{2} + \frac{13\lambda}{10}(1 - n\beta\mu_2)]n^2\alpha^3 l^2\sigma^2 + \frac{21\lambda n^2\beta^3 l^2\sigma^2}{10}.$$

46

By setting $\lambda = \frac{1}{10}, \varepsilon = \frac{3}{7}$ and $\eta = \frac{73l^2}{2\mu_2^2}$ and using $l^2 \geq \mu_1\mu_2$, we conclude:

$$C_1 \leq 1 - \mu_1 n\alpha/2,$$

$$C_2 \leq 1 - \frac{l^2 n\alpha}{2\mu_2} \leq 1 - \mu_1 n\alpha/2,$$

$$C_3 \leq 0.21 n^2 \alpha^3 l^2 \sigma^2 (3 + \eta^3)$$

From the above bounds, it follows that:

$$\mathbb{E}_k[V_\lambda(\mathbf{z}_0^{k+1})] \leq (1 - \mu_1 n\alpha/2)V_\lambda(\mathbf{z}_0^k) + 0.21 n^2 \alpha^3 l^2 \sigma^2 (3 + \eta^3).$$

By taking expectations with respect to the epoch iterates $(\mathbf{x}_0^j, \mathbf{y}_0^j)_{j=1}^k$ and applying the law of iterated expectation, we obtain the following:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{k+1})] \leq (1 - \mu_1 n\alpha/2)\mathbb{E}[V_\lambda(\mathbf{z}_0^k)] + 0.21 n^2 \alpha^3 l^2 \sigma^2 (3 + \eta^3).$$

By unrolling the above inequality for K steps, we obtain the following:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + 0.42 n\alpha^2 l^2 \sigma^2 (3 + \eta^3)/\mu_1. \tag{E.27}$$

By setting $\alpha = \min\{1/5\eta nl, 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/\mu_1 nK\}$ and using $\eta \leq 37\kappa^2$, we observe that:

$$n\alpha^2 l^2 \sigma^2 (3 + \eta^3)/\mu_1 \leq \frac{16 l^2 \sigma^2 (3 + \eta^3)}{\mu_1^3} \frac{\log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{nK^2}$$

$$\leq \frac{c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{\mu_1 nK^2},$$

where $c > 0$ is a constant that is independent of κ, μ and σ^2. By substituting the above into (E.27), we get:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + \frac{c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{\mu_1 nK^2}. \tag{E.28}$$

To complete the proof, we consider the following cases:

Case 1: $1/5\eta nl \leq 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/\mu_1 nK$

In this case, $\alpha = 1/5\eta nl$ and hence, we get:

$$(1 - n\alpha\mu_1/2)^K V_\lambda(\mathbf{z}_0) \leq e^{-n\alpha\mu_1 K/2}V_\lambda(\mathbf{z}_0) \leq e^{-\frac{K}{365\kappa^3}}V_\lambda(\mathbf{z}_0).$$

Substituting into (E.28), we get:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq e^{-\frac{K}{365\kappa^3}}V_\lambda(\mathbf{z}_0) + \frac{c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{\mu_1 nK^2}. \tag{E.29}$$

Case 2: $1/5\eta nl \leq 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/(\mu_1 nK)$

In this case, $\alpha = 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/\mu_1 nK$ and we conclude that,

$$(1 - n\alpha\mu_1/2)^K V_\lambda(\mathbf{z}_0) \leq e^{-n\alpha\mu_1 K/2}V_\lambda(\mathbf{z}_0) \leq e^{-2\log(V_\lambda(\mathbf{z}_0)^{1/2}n^{1/2}K)}V_\lambda(\mathbf{z}_0) \leq \frac{1}{nK^2}.$$

By substituting into (E.28), we get:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq \frac{\mu_1 + c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{\mu_1 nK^2}. \tag{E.30}$$

By taking the maximum of the right hand side of (E.29) and (E.30), we obtain the following rate, which holds for any $K \geq 1$:

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] \leq e^{-\frac{K}{365\kappa^3}}V_\lambda(\mathbf{z}_0) + \frac{\mu_1 + c\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{\mu_1 nK^2}.$$

Suppressing constant terms and logarithmic factors, we get,

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] = \tilde{O}(e^{-\frac{K}{365\kappa^3}} + 1/nK^2).$$

$\square$

47

Bounded iterate assumption   We now demonstrate that our analysis of AGDA-RR easily adapts to the bounded iterate setting. To this end, we assume that all the iterates of AGDA-RR, i.e., the iterates $(\mathbf{x}_i^k, \mathbf{y}_0^k)$ and $(\mathbf{x}_n^k, \mathbf{y}_i^k)$ where $i \in [n]$ and $k \in [K]$ are contained within a compact set $\mathcal{W}$. Furthermore, we note that the l-smoothness of the f_i's guarantees that the function $G_\sigma(\mathbf{x}, \mathbf{y}) = 1/n \sum_{i=1}^n |\omega_i(\mathbf{x}, \mathbf{y}) - \nu(\mathbf{x}, \mathbf{y})|^2$ is non-negative and continuous. Hence, $\sigma^2 = \max_{(\mathbf{x}, \mathbf{y}) \in \mathcal{W}} G_\sigma(\mathbf{x}, \mathbf{y})$ is non-negative and finite. Since $(\mathbf{x}_0^k, \mathbf{y}_0^k) \in \mathcal{W}$ and $(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \in \mathcal{W}$, we conclude that

$$1/n \sum_{i=1}^n \left|\omega_i(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nu(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 \leq \sigma^2,$$

$$1/n \sum_{i=1}^n \left|\omega_i(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) - \nu(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right|^2 \leq \sigma^2.$$

Substituting the above inequalities into Step 2 of Theorem E.1 shows that our convergence analysis of AGDA-RR straightforwardly adapts to the bounded iterate setting. The same argument also applies to our analysis of AGDA-AS.

Convergence in terms of dist$(\mathbf{z}, \mathcal{Z}^*)^2$ Let $\mathcal{Z}^*$ denote the set of all saddle points of the objective F and let dist$(\mathbf{z}, \mathcal{Z}^*)$ denote the distance of a point $\mathbf{z} \in \mathbb{R}^d$ from $\mathcal{Z}^*$. As we demonstrate in Appendix E.4, the Lyapunov function V_λ used in our analysis satisfies the following:

$$\text{dist}(\mathbf{z}, \mathcal{Z}^*)^2 \leq \max\left\{\frac{2}{\mu_1}(\frac{l^2}{2\mu_2} + 1), \frac{4}{\lambda \mu_2}\right\} V_\lambda(\mathbf{z}).$$

Using the above relation, the convergence guarantee of Theorem E.1 can be translated to a guarantee of the form $\mathbb{E}[\text{dist}(\mathbf{z}_0^{K+1}, \mathcal{Z}^*)^2] = \tilde{O}(e^{-K/365\kappa^3} + 1/nK^2)$. Thus, the epoch iterates $\mathbf{z}_0^{K+1}$ of AGDA-RR converge (in expectation) to some saddle point of F at a rate of $\tilde{O}(e^{-K/365\kappa^3} + 1/nK^2)$. A similar argument also applies to AGDA-AS.

E.3   Analysis of AGDA-AS

Theorem E.2 (Convergence of AGDA-AS). *Let Assumptions 1, 3, and 4 be satisfied and let $\eta = 73l^2/2\mu_2^2$. Then, for any $\alpha \leq 1/5\eta nl$, $\beta = \eta\alpha$, and $K \geq 1$, the iterates of AGDA-AS satisfy the following:*

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \leq (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + 2n^2\alpha^2 l^2 \sigma^2(2 + 3\eta^3)/\mu_1.$$

Setting $\alpha = \min\{1/5\eta nl, 4\log(V_\lambda(\mathbf{z}_0)K)/\mu_1 nK\}$ results in the following last iterate convergence guarantee which holds for any $K \geq 1$:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \leq e^{-K/365\kappa^3} V_\lambda(\mathbf{z}_0) + \frac{\mu_1 + \hat{c}\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)K)}{\mu_1 K^2}$$

$$= \tilde{O}(e^{-K/365\kappa^3} + 1/K^2),$$

where $\kappa = \max\{l/\mu_1, l/\mu_2\}$ and $\hat{c} > 0$ is a numerical constant that is independent of l, μ_1, μ_2 and σ^2.

Proof. **Step 1: Epoch level update rule** We highlight that the derivation of the epoch level update rule in Step 1 of Theorem E.1 applies to any permutations $\tau_k, \pi_k \in \mathbb{S}_n$. Thus, repeating the same arguments gives us the following epoch level update rule for AGDA-AS:

$$\mathbf{x}_0^{k+1} = \mathbf{x}_0^k - n\alpha \nabla_\mathbf{x} F(\mathbf{x}_0^k, \mathbf{y}_0^k) + \alpha^2 \mathbf{r}_k, \tag{E.31}$$

$$\mathbf{y}_0^{k+1} = \mathbf{y}_0^k + n\beta \nabla_\mathbf{y} F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) + \beta^2 \mathbf{s}_k, \tag{E.32}$$

where $\mathbf{r}_k$ and $\mathbf{s}_k$ are defined as follows:

$$\mathbf{r}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod}_{t=i+2}^n (\mathbf{I} - \alpha \mathbf{H}_{\tau_k(t)})\right] \mathbf{H}_{\tau_k(i+1)} \sum_{j=1}^i \nabla_\mathbf{x} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k), \tag{E.33}$$

48

$$\mathbf{s}_k = \sum_{i=1}^{n-1} \left[\widetilde{\prod_{t=i+2}^{n}} (\mathbf{I} + \beta \mathbf{J}_{\pi_k(t)}) \right] \mathbf{J}_{\pi_k(i+1)} \sum_{j=1}^{i} \nabla_{\mathbf{y}} f_{\pi_k(j)}(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k). \qquad (\text{E.34})$$

As before, $\mathbf{H}_{\tau_k(i)}$ and $\mathbf{J}_{\pi_k(i)}$ are defined as,

$$\mathbf{H}_{\tau_k(i)} = \int_0^1 \nabla_{\mathbf{xx}}^2 f_{\tau_k(i)}(\mathbf{x}_0^k + t(\mathbf{x}_{i-1}^k - \mathbf{x}_0^k), \mathbf{y}_0^k) \mathrm{d}t,$$

$$\mathbf{J}_{\pi_k(i)} = \int_0^1 \nabla_{\mathbf{yy}}^2 f_{\pi_k(i)}(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k + t(\mathbf{y}_{i-1}^k - \mathbf{y}_0^k)) \mathrm{d}t.$$

Step 2: Uniform bound on the noise terms Once again, we repeat the same computations as in Step 2 of Theorem E.1 to obtain:

$$|\mathbf{r}_k| \le le^{1/5} \left[n^2 \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| / 2 + \sum_{i=1}^{n-1} i |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - 1/i \sum_{j=1}^{i} \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)| \right].$$

Defining $d_{k,i} = |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - 1/i \sum_{j=1}^{i} \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|$ and using Young's inequality gives us the following:

$$|\mathbf{r}_k| \le le^{1/5} \left[n^2 \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| / 2 + \sum_{i=1}^{n-1} i d_{k,i} \right], \qquad (\text{E.35})$$

$$|\mathbf{r}_k|^2 \le 2l^2 e^{2/5} \left[n^4 \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 / 4 + (n-1) \sum_{i=1}^{n-1} i^2 d_{k,i}^2 \right]. \qquad (\text{E.36})$$

We proceed by using the bounded gradient variance assumption as follows:

$$\begin{aligned}
(n-1) \sum_{i=1}^{n-1} i^2 d_{k,i}^2 &= (n-1) \sum_{i=1}^{n-1} |\sum_{j=1}^{i} \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|^2 \\
&\le (n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{i} |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|^2 \\
&\le (n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{n} |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|^2 \\
&\le (n-1) \sum_{i=1}^{n-1} i \sum_{j=1}^{n} |\nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla_{\mathbf{x}} f_{\tau_k(j)}(\mathbf{x}_0^k, \mathbf{y}_0^k)|^2 \\
&\le (n-1)(n\sigma^2) \sum_{i=1}^{n-1} i \le n^4 \sigma^2 / 2.
\end{aligned}$$

Substituting the above into (E.36) gives us the following:

$$|\mathbf{r}_k|^2 \le \frac{3n^4 l^2}{4} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right|^2 + \frac{3\sigma^2 l^2 n^4}{2}. \qquad (\text{E.37})$$

Furthermore, by Young's inequality, we obtain,

$$\sum_{i=1}^{n-1} i d_{k,i} \le [(n-1) \sum_{i=1}^{n-1} i^2 d_{k,i}^2]^{1/2} \le n^2 \sigma / \sqrt{2}.$$

By substituting the above into (E.35), we obtain the following:

$$|\mathbf{r}_k| \le \frac{2n^2 l}{3} \left| \nabla_{\mathbf{x}} F(\mathbf{x}_0^k, \mathbf{y}_0^k) \right| + \sigma l n^2. \qquad (\text{E.38})$$

49

Performing similar computations for $\mathbf{s}_k$ gives us the following bounds:

$$|\mathbf{s}_k| \leq \frac{2n^2 l}{3}\left|\nabla_{\mathbf{y}}F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right| + \sigma l n^2, \tag{E.39}$$

$$|\mathbf{s}_k|^2 \leq \frac{3n^4 l^2}{4}\left|\nabla_{\mathbf{y}}F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)\right|^2 + \frac{3\sigma^2 l^2 n^4}{2}. \tag{E.40}$$

We highlight that the bounds (E.11), (E.12) (E.13) and (E.14) hold uniformly for any permutations $\tau_k, \pi_k \in \mathbb{S}_n$.

Step 3: Obtaining a convergence guarantee As before, we define $a_k = \Phi(\mathbf{x}_0^k) - \Phi^*$ and $b_k = \Phi(\mathbf{x}_0^k) - F(\mathbf{x}_0^k, \mathbf{y}_0^k)$ which implies that $V_\lambda(\mathbf{z}_0^k) = a_k + \lambda b_k$.

We follow the same procedure as Step 3 of Theorem E.1 and substitute the uniform bounds (E.11), (E.12) (E.13) and (E.14) wherever appropriate to obtain the following inequalities:

$$\Phi(\mathbf{x}_0^{k+1}) - \Phi(\mathbf{x}_0^k) \leq n\alpha\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 - \frac{n\alpha}{2}\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$
$$+ \frac{3n^3\alpha^3 l^2}{4}\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + \frac{3n^3\alpha^3 l^2 \sigma^2}{2}, \tag{E.41}$$

$$F(\mathbf{x}_0^k, \mathbf{y}_0^k) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k) \leq \left(\frac{3n\alpha}{2} + \frac{13}{10}ln^2\alpha^2 + \frac{3n^4\alpha^4 l^3}{8}\right)\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + \frac{9\alpha^3 n^3 l^2 \sigma^2}{10}, \tag{E.42}$$

$$\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^{k+1}) \leq (1 - n\beta\mu_2)[\Phi(\mathbf{x}_0^{k+1}) - F(\mathbf{x}_0^{k+1}, \mathbf{y}_0^k)] + \frac{23n^2\beta^3 l^2 \sigma^2}{10}. \tag{E.43}$$

Once again, following the same steps as Step 3 of Theorem E.1, we conclude that,

$$a_{k+1} \leq a_k + n\alpha\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 - \frac{n\alpha}{2}\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$
$$+ \frac{3n\alpha}{100}\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + \frac{3n^3\alpha^3 l^2 \sigma^2}{2}. \tag{E.44}$$

$$b_{k+1} \leq (1 - n\beta\mu_2)b_k + (1 - n\beta\mu_2)n\alpha\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k) - \nabla\Phi(\mathbf{x}_0^k)\right|^2 - (1 - n\beta\mu_2)\frac{n\alpha}{2}\left|\nabla\Phi(\mathbf{x}_0^k)\right|^2$$
$$+ (1 - n\beta\mu_2)\frac{9n\alpha}{5}\left|\nabla_{\mathbf{x}}F(\mathbf{x}_0^k, \mathbf{y}_0^k)\right|^2 + (1 - n\beta\mu_2)\frac{12n^3\alpha^3 l^2 \sigma^2}{5} + \frac{23n^3\beta^3 l^2 \sigma^2}{10}. \tag{E.45}$$

By setting $\lambda = \frac{1}{10}$ and $\eta = \frac{73l^2}{2\mu_2^2}$ and following the same steps as in Step 3 of Theorem E.1, we obtain the following update equation for $V_\lambda(\mathbf{z}_0^{k+1})$:

$$V_\lambda(\mathbf{z}_0^{k+1}) \leq (1 - \mu_1 n\alpha/2)V_\lambda(\mathbf{z}_0^{k+1}) + n^3\alpha^3 l^2 \sigma^2(2 + 3\eta^3).$$

By unrolling the above inequality for K steps we obtain the following:

$$V_\lambda(\mathbf{z}_0^{K+1}) \leq (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + 2n^2\alpha^2 l^2 \sigma^2(2 + 3\eta^3)/\mu_1.$$

Since the above holds for any sequence of permutations $\tau_1, \pi_1, \ldots \tau_K, \pi_K \in \mathbb{S}_n$, we conclude that:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \leq (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + 2n^2\alpha^2 l^2 \sigma^2(2 + 3\eta^3)/\mu_1. \tag{E.46}$$

Setting $\alpha = \min\{1/5\eta nl, 4\log(V_\lambda(\mathbf{z}_0)K)/\mu_1 nK\}$ and using $\eta \leq 37\kappa^2$, we observe that:

$$n^2\alpha^2 l^2 \sigma^2(2 + 3\eta^3)/\mu_1 \leq \frac{16l^2\sigma^2(2 + 3\eta^3)}{\mu_1^3}\frac{\log^2(V_\lambda(\mathbf{z}_0)n^{1/2}K)}{K^2}$$
$$\leq \frac{\hat{c}\kappa^8\sigma^2\log^2(V_\lambda(\mathbf{z}_0)K)}{\mu_1 K^2},$$

50

where $\hat{c} > 0$ is a numerical constant that is independent of κ, μ and σ^2. By substituting the above into (E.46), we get:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \le (1 - \mu_1 n\alpha/2)^K V_\lambda(\mathbf{z}_0) + \frac{\hat{c}\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)K)}{\mu_1 K^2}. \qquad \text{(E.47)}$$

Case 1: $1/5\eta n l \le 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/\mu_1 n K$

In this case, $\alpha = 1/5\eta n l$ and hence, we get:

$$(1 - n\alpha\mu_1/2)^K V_\lambda(\mathbf{z}_0) \le e^{-n\alpha\mu_1 K/2} V_\lambda(\mathbf{z}_0) \le e^{-\frac{K}{365\kappa^3}} V_\lambda(\mathbf{z}_0).$$

By substituting into (E.47), we get:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \le e^{-\frac{K}{365\kappa^3}} V_\lambda(\mathbf{z}_0) + \frac{\hat{c}\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)K)}{\mu_1 K^2}. \qquad \text{(E.48)}$$

Case 2: $1/5\eta n l \le 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/(\mu_1 n K)$

In this case, $\alpha = 4\log(V_\lambda(\mathbf{z}_0)n^{1/2}K)/\mu_1 n K$ and we conclude that,

$$(1 - n\alpha\mu_1/2)^K V_\lambda(\mathbf{z}_0) \le e^{-n\alpha\mu_1 K/2} V_\lambda(\mathbf{z}_0) \le e^{-2\log(V_\lambda(\mathbf{z}_0)^{1/2}n^{1/2}K)} V_\lambda(\mathbf{z}_0) \le \frac{1}{nK^2}.$$

By substituting into (E.47), we get:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \le \frac{\mu_1 + \hat{c}\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)K)}{\mu_1 K^2}. \qquad \text{(E.49)}$$

Taking the maximum of the right hand side of (E.48) and (E.49), we obtain the following rate, which holds for any $K \ge 1$:

$$\max_{\tau_1, \pi_1, \ldots, \tau_K, \pi_K \in \mathbb{S}_n} V_\lambda(\mathbf{z}_0^{K+1}) \le e^{-\frac{K}{365\kappa^3}} V_\lambda(\mathbf{z}_0) + \frac{\mu_1 + \hat{c}\kappa^8 \sigma^2 \log^2(V_\lambda(\mathbf{z}_0)K)}{\mu_1 K^2}.$$

Suppressing constant terms and logarithmic factors, we get,

$$\mathbb{E}[V_\lambda(\mathbf{z}_0^{K+1})] = \tilde{O}(e^{-\frac{K}{365\kappa^3}} + 1/K^2).$$

$\square$

E.4 Convergence of AGDA-RR and AGDA-AS to a Saddle Point

In this section, we demonstrate that our convergence guarantees for AGDA-RR and AGDA-AS can be expressed in terms of squared distance to the set of saddle points of F, denoted as $\mathcal{Z}^*$. We do so by upper bounding $\text{dist}(\mathbf{z}, \mathcal{Z}^*)^2$ in terms of $V_\lambda(\mathbf{z})$. We begin by establishing some basic properties of $\mathcal{Z}^*$ as follows.

Lemma E.5 (Set of Saddle Points of 2PŁ functions). *Let $F : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}^{d_\mathbf{y}} \to \mathbb{R}$ be an l-smooth function satisfying Assumption 3 (with constants μ_1 and μ_2) and let $\mathcal{Z}^*$ denote the set of saddle points of F. Then $\mathcal{Z}^*$ is a closed set and $\text{dist}(\mathbf{z}, \mathcal{Z}^*) = \min_{\mathbf{z}^* \in \mathcal{Z}^*} |\mathbf{z} - \mathbf{z}^*|$ is a well defined continuous function of $\mathbf{z}$. Furthermore, for any $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$, the set $\mathcal{Y}^*(\mathbf{x}) = \arg\max_{\hat{\mathbf{y}} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \hat{\mathbf{y}})$ is a closed set.*

Proof. By the l-smoothness of F, $\nabla_\mathbf{x} F(\mathbf{x}, \mathbf{y})$ and $\nabla_\mathbf{y} F(\mathbf{x}, \mathbf{y})$ are Lipschitz continuous functions, which implies that $\eta(\mathbf{z}) = [\nabla_\mathbf{x} F(\mathbf{z}), \nabla_\mathbf{y} F(\mathbf{z})]$ is a continuous map. Furthermore, by Lemma E.1, the set of saddle points of F are exactly the set of stationary points of F, which means,

$$\mathcal{Z}^* = \{\mathbf{z} \in \mathbb{R}^d \mid \nabla_\mathbf{x} F(\mathbf{z}) = 0, \nabla_\mathbf{y} F(\mathbf{z}) = 0\} = \eta^{-1}(\{0\}).$$

Hence, $\mathcal{Z}^*$ is a closed set as it is the inverse image of the closed set $\{0\}$ under a continuous map. Consequently, $\text{dist}(\mathbf{z}, \mathcal{Z}^*) = \min_{\mathbf{z}^* \in \mathcal{Z}^*} |\mathbf{z} - \mathbf{z}^*|$ is a well defined continuous function of $\mathbf{z}$.

Consider any $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$. Since F is a 2PŁ function, $-F(\mathbf{x}, .)$ satisfies the (one-sided) PŁ inequality. Thus, $\mathbf{y} \in \mathcal{Y}^*(\mathbf{x})$ if and only if $\nabla_\mathbf{y} F(\mathbf{x}, \mathbf{y}) = 0$, i.e., $\mathcal{Y}^*(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}} \mid \nabla_\mathbf{y} F(\mathbf{x}, \mathbf{y}) = 0\}$. By the continuity of $\nabla_\mathbf{y} F(\mathbf{x}, .)$, we conclude that $\mathcal{Y}^*(\mathbf{x})$ is a closed set. $\square$

51

The following lemma, which is key to establishing the relation between $V_\lambda(\mathbf{z})$ and $\mathrm{dist}(\mathbf{z}, \mathcal{Z}^*)$, is an adaptation of Lemma A.3 of Nouiehed et al. [40] for 2PŁ functions.

Lemma E.6 (Lipschitzness of $\mathcal{Y}^*(\mathbf{x})$). *Let $F : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}^{d_\mathbf{y}} \to \mathbb{R}$ be an l-smooth function satisfying Assumption 3 (with constants μ_1 and μ_2). For any $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$, let $\mathcal{Y}^*(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y})$. Then, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_\mathbf{x}}$ and $\mathbf{y}^*(\mathbf{x}_1) \in \mathcal{Y}^*(\mathbf{x}_1)$, there exists a $\mathbf{y}^*(\mathbf{x}_2) \in \mathcal{Y}^*(\mathbf{x}_2)$ such that,*

$$|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)| \leq \frac{l}{2\mu_2} |\mathbf{x}_1 - \mathbf{x}_2| .$$

Proof. Consider any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_\mathbf{x}}$ and $\mathbf{y}^*(\mathbf{x}_1) \in \mathcal{Y}^*(\mathbf{x}_1)$. Since F is 2PŁ with constants μ_1 and μ_2, $-F(\mathbf{x}, .)$ satisfies the PŁ inequality for any $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$ with constant μ_2. Furthermore, by Lemma E.5, $\mathcal{Y}^*(\mathbf{x})$ is a closed set for any $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$. Thus, by Lemma A.3 of Nouiehed et al. [40], there exists a $\mathbf{y}^*(\mathbf{x}_2) \in \mathcal{Y}^*(\mathbf{x}_2)$ satisfying $|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)| \leq \frac{l}{2\mu_2} |\mathbf{x}_1 - \mathbf{x}_2|$. $\qquad \square$

Equipped with the above lemmas, we can finally upper bound $\mathrm{dist}(\mathbf{z}, \mathcal{Z}^*)$ in terms of $V_\lambda(\mathbf{z})$.

Lemma E.7. *Let $F : \mathbb{R}^{d_\mathbf{x}} \times \mathbb{R}^{d_\mathbf{y}} \to \mathbb{R}$ be an l-smooth function satisfying Assumption 3 (with constants μ_1 and μ_2). Define $\Phi(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y})$, $\Phi^* = \min_{\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}} \Phi(\mathbf{x})$, $\mathcal{X}^* = \arg\min_{\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}} \Phi(\mathbf{x})$ and $\mathcal{Y}^*(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathbb{R}^{d_\mathbf{y}}} F(\mathbf{x}, \mathbf{y})$. Moreover, for any $\lambda > 0$, define $V_\lambda(\mathbf{x}, \mathbf{y}) = [(\Phi(\mathbf{x}) - \Phi^*) + \lambda(\Phi(\mathbf{x}) - F(\mathbf{x}, \mathbf{y})]$. Then the following holds for any $\mathbf{z} \in \mathbb{R}^d$:*

$$dist(\mathbf{z}, \mathcal{Z}^*)^2 \leq \max \left\{ \frac{2}{\mu_1}(\frac{l^2}{2\mu_2^2} + 1), \frac{4}{\lambda\mu_2} \right\} V_\lambda(\mathbf{z}).$$

Proof. Since the set of saddle points of F is non-empty and Φ is a PŁ function (with constant μ_1) as per Lemma E.2, we can use the same arguments as Lemma E.5 to show that $\mathcal{X}^*$ is a closed set.

Consider any $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Let $\mathbf{x}^*$ be the projection of $\mathbf{x}$ on the closed set $\mathcal{X}^*$ and let $\mathbf{y}^*(\mathbf{x})$ be the projection of $\mathbf{y}$ on the closed set (as per Lemma E.5) $\mathcal{Y}^*(\mathbf{x})$.

Applying Lemma E.3 to $\Phi(\mathbf{x})$, we obtain,

$$\Phi(\mathbf{x}) - \Phi^* \geq \frac{\mu_1}{2} |\mathbf{x} - \mathbf{x}^*|^2 . \tag{E.50}$$

Similarly, applying Lemma E.3 to $-F(\mathbf{x}, \mathbf{y})$, we obtain,

$$\Phi(\mathbf{x}) - F(\mathbf{x}, \mathbf{y}) \geq \frac{\mu_2}{2} |\mathbf{y} - \mathbf{y}^*(\mathbf{x})|^2 . \tag{E.51}$$

By Lemma E.6, there exists a $\mathbf{y}^* \in \mathcal{Y}^*(\mathbf{x}^*)$ satisfying,

$$|\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*| \leq \frac{l}{2\mu_2} |\mathbf{x} - \mathbf{x}^*| . \tag{E.52}$$

Let $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$. We note that since $\mathbf{x}^* \in \mathcal{X}^*$ and $\mathbf{y}^* \in \mathcal{Y}^*(\mathbf{x}^*)$, $\mathbf{z}^*$ is a global minimax point of F. Thus by Lemma E.1, $\mathbf{z}^*$ is a saddle point of F, i.e., $\mathbf{z}^* \in \mathcal{Z}^*$. We now proceed as follows:

$$|\mathbf{y} - \mathbf{y}^*| \leq |\mathbf{y} - \mathbf{y}^*(\mathbf{x})| + |\mathbf{y}^*(\mathbf{x}) - \mathbf{y}^*|$$

$$\leq |\mathbf{y} - \mathbf{y}^*(\mathbf{x})| + \frac{l}{2\mu_2} |\mathbf{x} - \mathbf{x}^*| .$$

Applying Young's Inequality, we obtain,

$$|\mathbf{y} - \mathbf{y}^*|^2 \leq 2 |\mathbf{y} - \mathbf{y}^*(\mathbf{x})|^2 + \frac{l^2}{2\mu_2^2} |\mathbf{x} - \mathbf{x}^*|^2 , \tag{E.53}$$

which implies that.

$$|\mathbf{z} - \mathbf{z}^*|^2 = |\mathbf{x} - \mathbf{x}^*|^2 + |\mathbf{y} - \mathbf{y}^*|^2$$

$$\leq 2 |\mathbf{y} - \mathbf{y}^*(\mathbf{x})|^2 + (\frac{l^2}{2\mu_2^2} + 1) |\mathbf{x} - \mathbf{x}^*|^2 .$$

52

Substituting (E.50) and (E.51) in the above inequality, we obtain,

$$|\mathbf{z} - \mathbf{z}^*|^2 \leq \frac{4}{\mu_2}[\Phi(\mathbf{x}) - F(\mathbf{x}, \mathbf{y})] + \frac{2}{\mu_1}(\frac{l^2}{2\mu_2^2} + 1)[\Phi(\mathbf{x}) - \Phi^*]$$

$$\leq \max\left\{\frac{2}{\mu_1}(\frac{l^2}{2\mu_2^2} + 1), \frac{4}{\lambda\mu_2}\right\}[(\Phi(\mathbf{x}) - \Phi^*) + \lambda(\Phi(\mathbf{x}) - F(\mathbf{x}, \mathbf{y}))]$$

$$\leq \max\left\{\frac{2}{\mu_1}(\frac{l^2}{2\mu_2^2} + 1), \frac{4}{\lambda\mu_2}\right\}V_\lambda(\mathbf{z}).$$

Since $\mathbf{z}^* \in \mathcal{Z}^*$, we conclude that,

$$\text{dist}(\mathbf{z}, \mathcal{Z}^*)^2 \leq \max\left\{\frac{2}{\mu_1}(\frac{l^2}{2\mu_2^2} + 1), \frac{4}{\lambda\mu_2}\right\}V_\lambda(\mathbf{z}). \tag{E.54}$$

$\square$

Using Lemma E.7 and E.1, we can conclude that AGDA-RR converges in terms of $\text{dist}(\mathbf{z}, \mathcal{Z}^*)^2$ as $\mathbb{E}[\text{dist}(\mathbf{z}_0^{K+1}, \mathcal{Z}^*)^2] = \tilde{O}(e^{-K/365\kappa^3} + 1/nK^2)$. Since $\mathcal{Z}^*$ is a closed set, this implies that the final epoch iterate $\mathbf{z}_0^{K+1}$ of AGDA-RR converges (in expectation) to some saddle point of F, with the rate $\tilde{O}(e^{-K/365\kappa^3} + 1/nK^2)$. Similarly, using Using Lemma E.7 and E.2, we conclude that the final epoch iterate $\mathbf{z}_0^{K+1}$ of AGDA-AS converges to some saddle point of F with the rate $\text{dist}(\mathbf{z}_0^{K+1}, \mathcal{Z}^*)^2 = \tilde{O}(e^{-K/365\kappa^3} + 1/K^2)$.

F Literature Review

While the convergence properties of SGD (and SGDA) with uniform sampling have been well studied for a long time, with matching upper and lower bounds of $\Theta(1/nK)$ for both strongly convex minimization and strongly convex-strongly concave minimax optimization [44, 13] (n is the number of components in the finite-sum objective and K is the number of epochs), advances in the theoretical understanding of without-replacement sampling for SGD are a fairly recent phenomenon. This is generally attributed to the unavailability of provably unbiased gradient estimates, leading to a significantly more complicated analysis that often requires sophisticated mathematical tools.

Initial progress was made by the pioneering work of Recht and Re [45] who proposed the *non-commutative AM-GM conjecture*, and proved that this conjecture implies the faster convergence of without-replacement SGD. Unfortunately, the conjecture was recently disproved by Lai and Lim [27]. Subsequent progress was made in Gürbüzbalaban et al. [21], where the authors establish an asymptotic $O(1/K^2)$ convergence rate for GD with Random Reshuffling (or RR) on functions of the form $F(x) = 1/n \sum_{i=1}^n f_i(x)$ where F is a strongly convex quadratic and the components f_i satisfy a Hessian Lipschitz assumption. For the same function class, Gürbüzbalaban et al. [22] established asymptotic $O(1/K^2)$ rates for Gradient Descent with Shuffle Once (SO) and Incremental Gradient (IG). The analysis in both these works relied on the following intuition, also proposed earlier in Nedic and Bertsekas [37]: *since sampling without replacement ensures that each component function is used exactly once within an epoch, the overall progress made by RR (and IG) over a complete epoch closely tracks that of full batch GD*. To this end, the key technique in both these works was to interpret the time evolution of the epoch iterates $\mathbf{x}_0^k$ as full batch GD with added noise, and to subsequently obtain a convergence rate by determining the (asymptotic) rate of decay for the noise using Chung's Lemma. This approach still remains a highly useful strategy for understanding SGD without replacement, and has served as a key insight for several future works such as Haochen and Sra [23], Ahn et al. [1]. In fact, formally extending this insight to minimax optimization (and more generally, to variational inequalities) in a manner that can simultaneously handle RR, SO and adversarial shuffling, is a cornerstone of our analysis.

The first nonasymptotic guarantees for RR, which also characterize the dependence of the convergence rate on n, were given by Haochen and Sra [23], who establish a convergence rate of $\tilde{O}(1/n^2K^2 + 1/K^3)$ assuming $K \geq C\kappa\log(nK)$ for some $C > 0$, for the case when F is a strongly convex quadratic and the components f_i are convex quadratics. Under the assumption that F satisfies the Polyak-Łojasiewicz inequality and the components f_i are smooth and Hessian Lipschitz, Haochen and Sra [23] also establish an $\tilde{O}(1/n^2K^2 + 1/K^3)$ rate assuming $K \geq C\kappa^2\log(nK)$. These rates imply the

53

provable superiority of RR when $K \geq \Omega(\sqrt{n})$. Subsequent advances were made by Nagaraj et al. [36], who used sophisticated Wasserstein coupling arguments to establish a rate of $O(1/nK^2)$ for RR which holds under the epoch requirement of $K \geq C\kappa^2 \log(nK)$, assuming F is strongly convex and f_i are smooth and Lipschitz. This rate demonstrates the provable improvement of RR over uniform sampling as soon as the epoch requirement is satisfied.

Development of lower bounds for SGD without replacement began with Safran and Shamir [48] where the authors establish a lower bound of $\Omega(1/n^2K^2 + 1/nK^3)$ for RR on strongly convex quadratic F with f_i being convex quadratics, assuming constant step-sizes. Under the same assumptions, Safran and Shamir [48] also establish an $\Omega(1/nK^2)$ lower bound for SO and an $\Omega(1/K^2)$ lower bound for IG. Further progress on lower bounds and their tightness was made in Rajput et al. [43] where the authors establish a $\tilde{O}(1/n^2K^2 + 1/nK^3)$ convergence rate for RR on strongly convex quadratic F with convex quadratic f_i. Rajput et al. [43] also develop a lower bound of $\Omega(1/nK^2)$ for strongly convex F with smooth f_i, thereby establishing the tightness of the upper bounds in Nagaraj et al. [36] modulo logarithmic factors.

Recently, the theory of SGD without replacement has been significantly advanced by the parallel works of Ahn et al. [1] and Mishchenko et al. [34]. For the case of Polyak-Łojasiewicz F with smooth (but not necessarily convex) f_i, Ahn et al. [1] show that RR converges at an optimal rate (modulo logarithmic factors) of $\tilde{O}(1/nK^2)$ when $K \geq C\kappa \log(n^{1/2}K)$. For the same problem class, Mishchenko et al. [34] obtain an anytime-valid near-optimal rate of $\tilde{O}(e^{-K/4\kappa} + 1/nK^2)$ under a bounded gradient variance assumption. This rate matches the optimal rate modulo logarithmic factors when $K \geq C\kappa \log(n^{1/2}K)$. Ahn et al. [1] also analyze SO for this problem class and derive $\tilde{O}(1/nK^2)$ rates assuming $K \geq C\kappa^2 \log(n^{1/2}K)$. The obtained rate matches the lower bound of Safran and Shamir [48] modulo logarithmic factors. Concurrently, Mishchenko et al. [34] present an analysis that covers both RR and SO for strongly convex F with smooth convex f_i and obtain a rate of $\tilde{O}(e^{-K/4\kappa} + 1/nK^2)$. For IG, Nguyen et al. [39] obtain a rate of $\tilde{O}(1/K^2)$ assuming strongly convex F, smooth f_i and an epoch requirement of $K \geq 12\kappa^2 \log(K)$. Moving beyond the PŁ class, the recent work of Li et al. [28] analyze RR for the case when f_i are smooth and F satisfies a KŁ inequality and, using variable step-sizes, establish an asymptotic convergence rate of $O(1/K^2)$ when the KŁ exponent lies in $[0, 1/2]$.

Recent works have also analyzed RR and SO in the federated minimization setting. Notably, Yun et al. [56] developed an extension of the techniques used in Ahn et al. [1] to analyze RR variants of minibatch and local SGD, while Mishchenko et al. [35] extended the approach of Mishchenko et al. [34] to develop federated analogs of RR and SO. Among other developments, Tran et al. [53] propose a variant of SGD with momentum that is amenable to theoretical analysis under without-replacement sampling. These developments are also complemented by the recent negative result of Safran and Shamir [49] which analyzes the condition number dependence of RR, and through an intricate lower bound analysis on convex quadratics, show that RR and SO do not significantly outperform uniform sampling unless K is larger than the condition number.

Despite the recent wave of developments on SGD without replacement, theoretical analysis of sampling without replacement for minimax optimization has remained relatively unexplored. To the best of our knowledge, Yu et al. [55] and Maheshwari et al. [33] are the only prior works in this domain. Both these works focus only on Random Reshuffling and are restricted to finite-sum minimax problems of the form $F(x, y) = 1/n \sum_{i=1}^{n} f_i(x, y)$ where each f_i is smooth, Lipschitz, convex-concave and has bounded domain. Under this setting, Yu et al. [55] proposes a stochastic proximal point method with RR that exhibits a rate of $\tilde{O}(1/\sqrt{nK})$ whereas Maheshwari et al. [33] proposes a zeroth-order optimistic gradient method with RR that exhibits a rate of $\tilde{O}(nd^2/K^{1/4})$, where d is the dimension of the domain. The analysis in both these works are based on an application of the Wasserstein coupling technique of Nagaraj et al. [36] (for the convex case) to smooth, Lipschitz convex-concave minimax optimization.

G Experimental Details

As described in Section 5 of the main text, we perform our benchmarks on finite-sum strongly convex-strongly concave quadratic minimax games. The objective F and the components f_i are given

54

by:

$$F(\mathbf{x}, \mathbf{y}) = 1/n \sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{x}^T \mathbf{B}\mathbf{y} - \frac{1}{2}\mathbf{y}^T \mathbf{C}\mathbf{y},$$

$$f_i(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}_i\mathbf{x} + \mathbf{x}^T \mathbf{B}_i\mathbf{y} - \frac{1}{2}\mathbf{y}^T \mathbf{C}_i\mathbf{y} - \mathbf{u}_i^T \mathbf{x} - \mathbf{v}_i^T \mathbf{y},$$

where $\mathbf{A}$ and $\mathbf{C}$ are positive definite and $\mathbf{u}_i$ and $\mathbf{v}_i$ are chosen such that $\sum_{i=1}^{n} \mathbf{u}_i = \sum_{i=1}^{n} \mathbf{v}_i = 0$. In all our experiments, we set $n = 100$ and dim $\mathbf{x}$ = dim $\mathbf{y}$ = 25. We randomly generate the components f_i such that 20 randomly selected components are strongly convex-strongly concave. To this end, the components $\mathbf{A}_i$ are selected as follows. We begin by sampling a random orthogonal matrix $\mathbf{O}_A$, and then generate two d-dimensional vectors $\mathbf{m}_A$ and δ_A, such that each component of $\mathbf{m}_A$ is uniformly sampled from $[\mu_A, L_A]$ whereas that of δ_A is uniformly sampled from $[\mu_\delta, L_\delta]$. Subsequently, for each $i \in [n]$, we set $\mathbf{A}_i = \mathbf{O}_A \Lambda_i \mathbf{O}_A^T$ where $\Lambda_i = \mathrm{diag}(-\delta_A)$ or 20 randomly selected indexes and $\Lambda_i = \mathrm{diag}(5\mathbf{m}_A/4 + \delta_A/4)$ for the other 80 components. A similar procedure is followed for generating the matrices $\mathbf{B}$ and $\mathbf{C}$. For the vectors $\mathbf{u}_i$, we generate a vector δ_u with components that are uniformly sampled from $[\mu_\delta, L_\delta]$. Then, for 20 randomly sampled indexes, we set $\mathbf{u}_i = -\delta_u$ and for the rest, we set $\mathbf{u}_i = \delta_u/4$. The vectors $\mathbf{v}_i$ are sampled in a similar fashion. We run each algorithm for $K = 100$ epochs using constant step-sizes of the form $\alpha = \gamma/n$ where γ is chosen independently for each algorithm via grid search. In all our experiments, we set $\mu_A = \mu_C = 0.5$, $\mu_B = 5.0$ and $\mu_\delta = 50.0$. Furthermore, we set $L_A = L_C = 2\mu_A$, $L_B = 2\mu_B$ and $L_\delta = 2\mu_\delta$. We perform our experiments on a Jupyter notebook with a Python 3.7 kernel, executed on a 2.8 GHz Intel Core i7 processor with 8 GB of memory. The total running time is 40 minutes.

55