Stochastic Extragradient with Random Reshuffling: Improved Convergence for Variational Inequalities

Konstantinos Emmanouilidis
CS & MINDS
Johns Hopkins University

René Vidal
ESE, Radiology & IDEAS
University of Pennsylvania

Nicolas Loizou
AMS & MINDS
Johns Hopkins University

Abstract

The Stochastic Extragradient (SEG) method is one of the most popular algorithms for solving finite-sum min-max optimization and variational inequality problems (VIPs) appearing in various machine learning tasks. However, existing convergence analyses of SEG focus on its with-replacement variants, while practical implementations of the method randomly reshuffle components and sequentially use them. Unlike the well-studied with-replacement variants, SEG with Random Reshuffling (SEG-RR) lacks established theoretical guarantees. In this work, we provide a convergence analysis of SEG-RR for three classes of VIPs: (i) strongly monotone, (ii) affine, and (iii) monotone. We derive conditions under which SEG-RR achieves a faster convergence rate than the uniform with-replacement sampling SEG. In the monotone setting, our analysis of SEG-RR guarantees convergence to an arbitrary accuracy without large batch sizes, a strong requirement needed in the classical with-replacement SEG. As a byproduct of our results, we provide convergence guarantees for Shuffle Once SEG (shuffles the data only at the beginning of the algorithm) and the Incremental Extragradient (does not shuffle the data). We supplement our analysis with experiments validating empirically the superior performance of SEG-RR over the classical with-replacement sampling SEG.

1 Introduction

Minimax optimization and, more generally, variational inequality problems (VIPs) have received much attention in recent years, especially in the machine learning community. Several machine learning tasks, including Generative Adversarial Networks (GANs) [Arjovsky et al., 2017, Goodfellow et al., 2014], adversarial training of neural networks [Madry et al., 2018, Wang et al., 2021], reinforcement learning [Brown et al., 2020, Sokota et al., 2023], and distributionally robust learning [Namkoong and Duchi, 2016, Yu et al., 2022] are formulated as finite-sum min-max optimization problems,

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y) = \frac{1}{n}\sum_{i=1}^{n} f_i(x,y)\,, \tag{1}$$

with the goal of finding a solution $z^* = (x^*, y^*)^\top$ such that $f(x^*, y) \le f(x^*, y^*) \le f(x, y^*), \forall x \in \mathbb{R}^{d_1}, y \in \mathbb{R}^{d_2}$.

In this work, we focus on a more abstract formulation of problem (1), and we analyze algorithms for solving the following unconstrained finite-sum variational inequality problem (VIP): find $z^* \in \mathbb{R}^d$ such that

$$F(z^*) = \frac{1}{n}\sum_{i=1}^{n} F_i(z^*) = 0, \tag{2}$$

where $F_i : \mathbb{R}^d \to \mathbb{R}^d, \forall i \in [n], d = d_1 + d_2$. We denote with $\mathcal{Z}_* \subset \mathbb{R}^d$ the solution set of (2).

Problem (2) is quite general and covers a wide range of possible problem formulations. For example, when the operator $F(x)$ is the gradient of a convex function $f(x)$, then problem (2) is equivalent to the minimization of the function $f(x)$. In addition, if the min-max optimization problem (1) has convex-concave continuously differentiable f, then using the first-order optimality conditions it can be cast as a special case of (2) with $z = (x^\top, y^\top)^\top \in \mathbb{R}^d$ and $F(z) = (\nabla_x f(x,y)^\top, -\nabla_y f(x,y)^\top)^\top$.

In the typical large-scale regime of machine learning applications (n in problem (2) is large), stochastic iterative algorithms are preferred mainly because of their

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

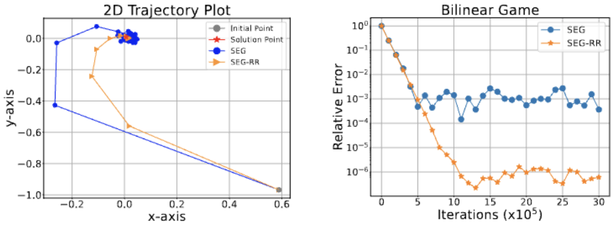

Figure 1: Bilinear Game. Left plot: 2D trajectory plot. Right plot: Relative error $\frac{\|z^k - z^*\|^2}{\|z^0 - z^*\|^2}$ as a function of the number of iterations.

cheap per-iteration cost. In that setting, we only assume to have access to a stochastic estimate of the operator F. Several papers have been devoted to the understanding and convergence analysis of stochastic variants of popular algorithms like the gradient method [Chen and Rockafellar, 1997], extragradient method [Gorbunov et al., 2022b, Korpelevich, 1976], and optimistic method [Gorbunov et al., 2022c, Popov, 1980]. Some recent works in the area include Beznosikov et al. [2023], Loizou et al. [2021] on the analysis of stochastic gradient descent ascent (SGDA), Gorbunov et al. [2022a], Hsieh et al. [2020], Mishchenko et al. [2020b] for SEG and Choudhury et al. [2023], Hsieh et al. [2019] for stochastic past extragradient methods.

Most existing analyses of stochastic algorithms for solving (2) focus on algorithms that use with-replacement sampling in their update rule. Specifically, a component F_i (or a minibatch) of the finite-sum structure of (2) is selected uniformly[1] at random in each step. However, most practical implementations of algorithms for solving finite-sum min-max problems and VIPs use without-replacement sampling, creating a gap between practical and theoretically understood approaches.

In the well-studied problem of solving finite-sum minimization problems (i.e., $\min_x \frac{1}{n} \sum f_i(x)$), practitioners prefer running popular algorithms that use without-replacement sampling. This is due to the remarkable ease of use and the better empirical performance compared to with-replacement variants [Bottou, 2009, 2012]. Unfortunately, the fact that the selected samples in an epoch of a without-replacement sampling algorithm are not independent of each other makes the analysis of the method more challenging. However, in the last few years, several works in the optimization literature were able to prove a faster convergence rate of SGD without-replacement under different scenarios [Ahn et al., 2020, Cai et al., 2023, Gürbüzbalaban et al., 2021, Mishchenko et al., 2020a, Nguyen et al., 2021, Safran and Shamir, 2020].

Despite the extensive use of without-replacement sampling, perhaps surprisingly, not many works focus on providing convergence guarantees for without-replacement sampling algorithms for solving min-max optimization problems and VIPs. Das et al. [2022] provide convergence guarantees for SGDA and the proximal point method (PPM) with without-replacement sampling for solving smooth and strongly convex-strongly concave problems satisfying a two-sided Polyak-Lojasiewicz inequality and show faster convergence, while Cho and Yun [2023] provide theoretical guarantees for SGDA with shuffling for solving structured non-monotone minimax problems. However, it is well known that SGDA fails to converge in simple monotone min-max problems (e.g., bilinear), while the PPM serves only as an implicit method.

The Stochastic Extragradient (SEG) method is one of the most popular algorithms for tackling finite-sum VIPs. The algorithm consists of two steps: a) an *extrapolation step* that computes a gradient update at the current iterate, and b) an *update step* that updates the current iterate using the value of the vector field at the extrapolation point. SEG comes in different forms [Gorbunov et al., 2022a]. One of the most common choices is same-sample SEG (S-SEG) given in the following update rule:

$$z^{k+1} = z^k - \gamma_1 F_i \left(z^k - \gamma_2 F_i(z^k) \right), \qquad \text{(S-SEG)}$$

where in each iteration, the same component $i \in [n]$ is sampled uniformly at random and used for the extrapolation (computation of $z^k - \gamma_2 F_i(z^k)$) and update (computation of z^{k+1}) steps. Existing works focusing on the convergence guarantees of SEG studied only with-replacement sampling strategies, similar to S-SEG. However, most practical implementations of SEG use without-replacement sampling.

A popular in practice but theoretically elusive update rule, belonging to the class of without-replacement sampling SEG, is *SEG with Random Reshuffling* given in SEG-RR (see also Algorithm 1). This is the method we pay most attention to in this work, as reflected in the title. In each epoch k, SEG-RR samples uniformly at random a permutation $\pi^k = \{\pi_0^k, \pi_1^k, \ldots, \pi_{n-1}^k\}$ of $[n] \coloneqq \{1, 2, \ldots, n\}$, and proceeds with n iterates of the form:

$$z_{i+1}^k = z_i^k - \gamma_1 F_{\pi_i^k} \left(z_i^k - \gamma_2 F_{\pi_i^k}(z_i^k) \right), \qquad \text{(SEG-RR)}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ are the step sizes in the update and extrapolation steps of the method, respectively. We then set $z_0^{k+1} = z_n^k$ and repeat the process for a total of K epochs. In SEG-RR (Alg. 1), a new permutation/shuffling is generated at the beginning of each epoch, which justifies the algorithm's name.

[1] Different with-replacement samplings can be used. Here, we use uniform distribution for ease of exposition.

Algorithm 1 SEG-RR

```
Given: $z_0$, step sizes $\gamma_1, \gamma_2$, number of epochs $K$
Initialize: $z_0^0 = z_0$
for all $k = 0, ..., K-1$ do
    Sample uniformly at random a permutation $\pi^k$ of $[n]$
    for all $i = 0, ..., n-1$ do
        $\bar{z}_i^k = z_i^k - \gamma_2 F_{\pi_i^k}(z_i^k)$
        $z_{i+1}^k = z_i^k - \gamma_1 F_{\pi_i^k}(\bar{z}_i^k)$
    end for
    $z_0^{k+1} = z_n^k$
end for
```

Algorithm 2 SEG-SO

```
Given: $z_0$, step sizes $\gamma_1, \gamma_2$, number of epochs $K$
Initialize: $z_0^0 = z_0$
Sample uniformly at random a permutation $\pi$ of $[n]$
for all $k = 0, ..., K-1$ do
    for all $i = 0, ..., n-1$ do
        $\bar{z}_i^k = z_i^k - \gamma_2 F_{\pi_i}(z_i^k)$
        $z_{i+1}^k = z_i^k - \gamma_1 F_{\pi_i}(\bar{z}_i^k)$
    end for
    $z_0^{k+1} = z_n^k$
end for
```

Algorithm 3 IEG

```
Given: $z_0$, step sizes $\gamma_1, \gamma_2$, number of epochs $K$
Initialize: $z_0^0 = z_0$
$\pi = [n]$ // maintain the order of the dataset
for all $k = 0, ..., K-1$ do
    for all $i = 0, ..., n-1$ do
        $\bar{z}_i^k = z_i^k - \gamma_2 F_{\pi_i}(z_i^k)$
        $z_{i+1}^k = z_i^k - \gamma_1 F_{\pi_i}(\bar{z}_i^k)$
    end for
    $z_0^{k+1} = z_n^k$
end for
```

Figure 2: Three variants of without-replacement sampling SEG (SEG-RR, SEG-SO, IEG)

As a proof of concept, in Figure 1, we compare the above two variants of SEG: S-SEG (with-replacement) and SEG-RR (without-replacement) on solving a simple two-dimension bilinear problem of the form (1), where we choose x and y to be scalars. As we can see in Fig. 1, SEG-RR converges to a smaller neighborhood of the min-max solution. Interestingly and perhaps surprisingly, in the left plot of Fig. 1, where we look at the trajectory of the two methods, the variant with random reshuffling (SEG-RR) reduces the rotation around the solution, which might explain its preference in practical implementations over the uniform sampling variant (S-SEG). This motivates us to study further the convergence guarantees of SEG-RR in different classes of problems. Our work aims to bridge the gap between the theoretical analysis and practical implementation of SEG by studying the following question:

Can Random Reshufling lead to improved theoretical and practical convergence for SEG in finite-sum VIPs?

1.1 Preliminaries

In this work, we assume that the operators F_i of problem (2) are L_i-Lipschitz. This implies that the operator F is also Lipschitz, and we will indicate with L its value. Throughout this work, we focus on three classes of operators F: (i) strongly monotone, (ii) affine, and (iii) monotone. Let us provide below the main definitions.

Definition 1.1 ($L-$Lipschitz)**.** An operator $F: \mathbb{R}^d \to \mathbb{R}^d$ is $L-$Lipschitz if there is $L > 0$:

$$\|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|, \quad \forall z_1, z_2 \in \mathbb{R}^d \quad (3)$$

We denote with $L_{max} = \max_{i \in [n]} L_i$, the maximum Lipschitz constant of the F_i operators in problem (2)

Definition 1.2 (Strongly monotone / monotone operator)**.** We say that an operator F is $\mu-$strongly monotone if there exist $\mu > 0$ such that $\forall z_1, z_2 \in \mathbb{R}^d$, $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu\|z_2 - z_2\|^2$. If $\mu = 0$, then $\forall z_1, z_2 \in \mathbb{R}^d, \langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq 0$, and we say that the operator is monotone.

Lastly, we also focus on the class of affine operators, a subclass of monotone VIPs, that can be seen as a generalization of bilinear min-max problems.

Definition 1.3 (Affine)**.** An operator $F: \mathbb{R}^d \to \mathbb{R}^d$ is affine if it there exist $Q \in \mathbb{R}^{d \times d}, b \in \mathbb{R}^d$ such that $F(z) = Qz + b$.

We denote with $\lambda_{min}^+(Q)$ the minimum non-zero eigenvalue of an affine and monotone operator F in problem (2).

On bounded variance. In the convergence analysis of SEG-RR, we do not assume bounded variance of the stochastic oracles F_i, i.e. there exists $c > 0$ such that $\mathbb{E}[\|\nabla F_i(z) - \nabla F(z)\|^2] \leq c, \forall z \in \mathbb{R}^d$, or growth conditions, i.e. there exist $c_1, c_2 > 0 : \mathbb{E}[\|\nabla F_i(z)\|^2] \leq c_1\|\nabla F(z)\|^2 + c_2$, $\forall z \in \mathbb{R}^d$. These conditions are typically assumed in the theoretical analysis of stochastic methods for solving finite-sum VIPs of the form (2), as they simplify the proofs [Juditsky et al., 2011, Lin et al., 2020a,b, Mishchenko et al., 2020b]. However, these assumptions are true only for a restrictive set of problems, and for large common classes of problems (e.g., unconstrained strongly monotone VIPs), they might not be even satisfied [Loizou et al., 2021]. Instead, we follow a recent line of work that uses the Lipschitz assumption to provide closed-form expressions for the upper bound on the variance [Choudhury et al., 2023, Gorbunov et al., 2022a, Loizou et al., 2021]. More specifically, in Appendix A we prove that if each F_i is L_i-Lipschitz, then the following bound on the variance holds: $\frac{1}{n}\sum_{i=0}^{n-1}\|F_i(z) - F(z)\|^2 \leq A\|z - z_*\|^2 + 2\sigma_*^2$

Algorithm	Citation	Strongly Monotone	Affine & Monotone	Monotone
SGDA	[Loizou et al., 2021]	$\tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$	✗	✗
SGDA-RR	[Das et al., 2022]	$\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$	✗	✗
S-SEG	[Gorbunov et al., 2022b, Hsieh et al., 2020]	$\tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$	✗$^{(*)}$
SEG-RR	This paper	$\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$	$\mathcal{O}\left(\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right)$

Table 1: Iteration Complexity of SEG with uniform with-replacement sampling S-SEG and SEG-RR after a certain number of epochs K that depends on the problem parameters. The $\tilde{\mathcal{O}}(\cdot)$ notation suppresses constant and logarithmic factors. $(*)$: In the monotone case, SEG with constant step sizes requires large batch sizes to achieve a given target accuracy $\epsilon > 0$.

where $A = \frac{2}{n}\sum_{i=0}^{n-1} L_i^2$ and $\sigma_*^2 = \frac{1}{n}\sum_{i=0}^{n-1}\|F_i(z_*)\|^2$. This new upper bound allows one to avoid the necessity of introducing any extra assumptions on the variance of the stochastic operators in the proofs.

1.2 Main Contributions

Our main contributions are summarized below. See also Table 1 for a comparison of iteration complexities of our results with closely related works.

1. **Strongly monotone or affine VIPs.** We prove the first convergence guarantees for SEG-RR for solving strongly monotone and affine VIPs. We show a linear convergence to a neighborhood of z_* when constant step sizes γ_1 and γ_2 are used, and we explain why a double stepsize selection is needed. In particular, in our theorems, we require the extrapolation stepsize to be larger than the update stepsize ($\gamma_2 > \gamma_1$), which aligns with recent results on the convergence of S-SEG [Gorbunov et al., 2022a, Hsieh et al., 2020]. In both strongly monotone and affine regimes, we prove improved convergence of random reshuffling over uniform with-replacement sampling by showing that after a certain number of epochs K, SEG-RR achieves an iteration complexity of $\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$ outperforming the $\tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$ iteration complexity of S-SEG. In the strongly monotone regime, this coincides with the benefit that SGDA-RR has over SGDA with uniform sampling proved in prior works. However, SGDA and SGDA-RR fail to converge to simple problems captured under the affine setting (e.g., bilinear minimax problems).

2. **Monotone VIPs: Convergence without large batch sizes.** In the monotone case, we prove a sublinear convergence of the weighted average iterate to a neighborhood around the solution. In particular, we prove that SEG-RR can reduce the neighborhood and reach any target accuracy ϵ by choosing appropriately the step sizes γ_1 and γ_2 of the method, establishing in this way an iteration complexity of $\mathcal{O}\left(\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right)$ after a certain number of epochs. This comes in stark contrast with the well-known results on the convergence of S-SEG for monotone problems, which require the use of large batch sizes, when constant step sizes are used in order to be able to reduce the neighborhood of convergence and reach any specific target accuracy $\epsilon > 0$.

3. **Further Convergence Guarantees: Other without-replacement samplings and novel stepsize selection.** As a byproduct of our analysis, we also provide convergence guarantees for two other popular without-replacements sampling variants of SEG, the Shuffle Once SEG (SEG-SO) (see Alg. 2), which shuffles the data only at the beginning of the algorithm, and the Incremental Extragradient (IEG) (see Alg. 3), which does not shuffle the data and processes them in the order that they appear in the dataset. For solving strongly monotone and affine VIPs, we also provide convergence guarantees under different stepsize rules. In particular, using a carefully constructed switching stepsize-rule, we prove a $O(1/k)$ rate to the exact solution. The suggested switching stepsize rule describes when one should switch from a constant to a decreasing stepsize regime, and it is the first time used in the analysis of algorithms utilizing without-replacement samplings. The details for these results are included in Appendix C.

4. **Numerical Experiments.** We show the benefits of SEG-RR by performing numerical experiments on finite-sum strongly-monotone quadratic and bilinear minimax problems, as well as on Wasserstein GANs for learning the mean of a multivariate Gaussian distribution. Our numerical findings corroborate our theoretical results.

2 Convergence Analysis

Let us now present our main theoretical results. We start by presenting a sketch of the proof techniques used in our theorems and explaining the difference/main challenge compared to the classical analysis of SEG. We, then, focus on the convergence guarantees for SEG-RR in three different classes of VIPs: strongly monotone, affine, and monotone.

2.1 Overview of Proof Techniques

The main challenge in the proof of SEG-RR compared to the one of S-SEG is that the stochastic oracles $F_{\pi_i^k}$ are no longer unbiased estimators of the deterministic operator F. Our proof is based on the key insight from previous works on random reshuffling in minimization problems [Ahn et al., 2020, Gürbüzbalaban et al., 2021, Haochen and Sra, 2019] that, for small enough step sizes, the epoch iterates z_0^k of a stochastic algorithm using without-replacement sampling approximately follow the trajectory of the same full-batch algorithm.

Building on the aforementioned idea, we manage to upper bound the distance to z^*, $\|z_0^{k+1}-z^*\|^2$, by three terms:

$$\|z_0^{k+1}-z_*\|^2 \leq C_1 \underbrace{\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2}_{T_1} + C_2 \underbrace{\sum_{i=0}^{n-1} \|F_{\pi_i^k}(z_i^k) - F(z_0^k)\|^2}_{T_2} + C_3 \underbrace{\sum_{i=0}^{n-1} \left\|z_i^k - z_0^k\right\|^2}_{T_3}, \qquad (4)$$

where $\hat{z}_0^k = z_0^k - \gamma_2 F(z_0^k)$ and C_1, C_2, C_3 are constants depending on the properties of the problem in hand.

The term T_1 in (4) can be interpreted as the distance of one step of the full-batch SEG algorithm, starting from the point z_0^k, to the optimum z_*. Thus, it serves as a measure of the progress that the algorithm that uses the full operator F makes. The term T_2, on the other hand, accounts for the fact that SEG-RR has access only to a stochastic oracle $F_{\pi_i^k}$ (not the full-batch operator F) per iteration. Using, in addition, the intuition that for small enough step sizes, the iterates z_i^k inside an epoch stay "close" to the initial point z_0^k, the second term T_2 measures the distance between the stochastic oracle $F_{\pi_i^k}$ from the (full-batch) operator F (though at different points). Lastly, the term T_3 indicates how "far" the points inside an epoch are from z_0^k. In our proofs, bounding each one of the three terms T_1, T_2, T_3 enables us to bound the distance of the iterate z_0^{k+1} from the optimum z_* and thus derive convergence guarantees for SEG-RR for the three different classes of VIPs under study.

2.2 Strongly Monotone VIPs

We focus on SEG-RR with constant step sizes γ_1 and γ_2. We first prove linear convergence to a neighborhood of the solution z_*. If, in addition, the total number of epochs K is available, we suggest a constant stepsize selection that depends on K, which allows us to prove that after a certain number of epochs K, SEG-RR achieves an iteration complexity of $\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$ outperforming the $\tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$ iteration complexity of S-SEG.

Theorem 2.1. Suppose that the operator F is μ-strongly monotone and each $F_i, \forall i \in [n]$ is L_i-Lipschitz.

1. Then the iterates of SEG-RR with constant step sizes $\gamma_2 = 2\gamma_1$, $\gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$ satisfy:

$$\mathbb{E}\left[\|z_0^k - z_*\|^2\right] \leq \left(1 - \frac{\gamma_1 n \mu}{4}\right)^k \|z_0 - z_*\|^2 + \frac{96L_{max}^2}{\mu^2}\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2 \qquad (5)$$

2. Let K be the total number of epochs SEG-RR is run. For step sizes satisfying $\gamma_2 = 2\gamma_1$ and $\gamma_1 = \min\left\{\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}, \frac{4\log(n^{1/2}K)}{\mu n K}\right\}$, the following holds:

$$\mathbb{E}\left[\|z_0^K - z_*\|^2\right] \leq \tilde{\mathcal{O}}\left(e^{-\frac{\mu^2 K}{L_{max}^2}} + \frac{1}{nK^2}\right) \qquad (6)$$

Theorem 2.1 indicates in inequality (5) that for constant step sizes the SEG-RR algorithm converges linearly to a neighborhood of the solution z_*. The neighborhood of convergence is proportional to the step sizes and the variance σ_*^2 at the optimum point. In particular, Theorem 2.1 indicates that the neighborhood around the solution z_* diminishes in relation to the step sizes of the algorithm as $\mathcal{O}(\gamma_1^2+\gamma_2^2)$. In comparison, S-SEG converges linearly to a neighborhood around the solution, with the neighborhood decreasing as $\mathcal{O}(\gamma_1+\gamma_2)$ (Theorem 3.1 of Gorbunov et al. [2022a]). Thus, although both algorithms achieve a linear convergence rate to a neighborhood around the solution, the without-replacement sampling variant will converge for the same step sizes to a smaller neighborhood of z_*.

In addition, with the total number of epochs K available, inequality (6) of Theorem 2.2 establishes the iteration complexity of SEG-RR for achieving an error $\mathbb{E}\left[\|z_0^K - z_*\|^2\right] \leq \epsilon$. More specifically, after a certain number of epochs satisfying $K \geq \kappa^2 \log(nK^2)$, where $\kappa = \frac{L_{max}}{\mu}$ is the condition number, the second term dominates in the iteration complexity and thus $\mathbb{E}\left[\|z_0^K - z_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$. In contrast, the iteration complexity of S-SEG in Gorbunov et al. [2022a] is $\tilde{\mathcal{O}}\left(e^{-\frac{\mu n K}{L_{max}}} + \frac{1}{nK}\right)$ and thus after the same number of epochs $K \geq \kappa^2 \log(nK^2)$ the distance from the solution is $\mathbb{E}\left[\|z_0^K - z_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$. In this case, SEG-RR will require less number of epochs (equivalently iterations) to achieve an accuracy ϵ. The difference in the iteration complexity of SEG-RR and S-SEG showcases the benefit of random reshuffling over uniform with-replacement sampling.

We note, also, that in the strongly monotone setting the iteration complexity of S-SEG with step sizes $\gamma_1 = \frac{1}{6L_{max}}, \gamma_2 = 4\gamma_1$, as shown in Gorbunov et al. [2022a], depends on the condition number as $\mathcal{O}\left(e^{-\kappa}\right)$. Despite this is a better dependence than the one in Theorem 2.1, we highlight that for the step sizes of Theorem 2.1 both SEG-RR and S-SEG have the same $\mathcal{O}(e^{-\kappa^2})$ dependence. Hence, for the step sizes of Theorem 2.1, SEG-RR and S-SEG will converge with the same rates to the corresponding neighborhoods of solution. Proving convergence of random reshuffling with larger step sizes and better dependence on the condition number is still an open problem.

We, lastly, compare our results on the convergence of SEG-RR with the SGDA-RR algorithm, which is the other frequently used method for solving strongly monotone problems. The iteration complexity of SGDA-RR for step sizes $\gamma_1 = \mathcal{O}\left(\frac{\log(n^{\frac{1}{2}}K^2)}{nK}\right)$, where K is the total number of epochs the algorithm is run, is $\tilde{\mathcal{O}}\left(e^{-\frac{\mu^2 K}{5L^2}} + \frac{1}{nK^2}\right)$, as established in Das et al. [2022]. In our Theorem 2.1, we establish the same iteration complexity (up to constant factors) with SGDA-RR. However, SEG-RR is able to solve VIPs beyond the strongly monotone regime, which is not the case of SGDA-RR, and we examine that below.

2.3 Affine VIP Operators

We, now, consider the setting where the variational inequality operator is affine and has the following form:

$$F(z) = \frac{1}{n}\sum_{i=0}^{n-1} Q_i z + b_i \tag{7}$$

where $F(z)$ has a finite-sum structure with each $F_i(z) = Q_i z + b_i$. This setting serves as a generalization of any bilinear min-max optimization problem. For more details on the connection of bilinear games with affine variational inequalities, we refer the interested reader to Appendix A.3.

Similarly to the strongly monotone regime, we focus on the convergence of SEG-RR with constant step sizes. We prove linear convergence to a neighborhood of the solution z_*. If, in addition, the total number of epochs K is available, we show that after a certain number of epochs K, SEG-RR achieves an iteration complexity of $\tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$. In this setting, since there might be multiple solutions z_*, we use as measure of convergence the $\mathrm{dist}(z_0^k, \mathcal{Z}_*) = \min_{z_* \in \mathcal{Z}_*} \|z_k^0 - z_*\|^2$, which is the distance of the iterate z_0^k from the solution set $\mathcal{Z}_*$.

Theorem 2.2. Suppose that each F_i, $\forall i \in [n]$, is monotone, affine and L_i-Lipschitz.

1. Then the iterates of SEG-RR with step sizes $\gamma_2 = 4\gamma_1$, $\gamma_1 \leq \frac{\lambda^+_{min}(Q)}{2\sqrt{120n}L^2_{max}}$ satisfy:

$$\mathbb{E}\left[\mathrm{dist}(z_0^k, \mathcal{Z}_*)\right] \leq \left(1 - \frac{\gamma_1 n \lambda^+_{min}(Q)}{2}\right)^k \mathrm{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}\left[4n(n+25)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2}{\lambda^+_{min}(Q)^2 n^2} \tag{8}$$

2. Let K be the total number of epochs the SEG-RR is run. For step sizes satisfying $\gamma_2 = 4\gamma_1$, $\gamma_1 \leq \min\left\{\frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}}, \frac{2\log(n^{1/2}K)}{\lambda^+_{min}(Q)nK}\right\}$, it holds:

$$\mathbb{E}\left[\mathrm{dist}(z_0^K, \mathcal{Z}_*)\right] \leq \tilde{\mathcal{O}}\left(e^{\frac{K\lambda^+_{min}{}^2(Q)}{4\sqrt{120}L^2_{max}}} + \frac{1}{nK^2}\right) \tag{9}$$

Theorem 2.2 indicates in (8) that SEG-RR achieves a linear convergence to a neighborhood of the solution z_*, which is proportional to the step sizes and the variance σ_*^2 at the optimum point. We highlight that the neighborhood of convergence decreases as $\mathcal{O}(\gamma_1^2 + \gamma_2^2)$. In contrast, Hsieh et al. [2020] establish for S-SEG a linear rate to a neihgbourhood that decreases as $\mathcal{O}(\gamma_1 + \gamma_2)$. For the step sizes suggested in Theorem 2.2, both S-SEG and SEG-RR converge with a linear rate, however, SEG-RR converges to a smaller neighborhood around the solution z_*.

The second point in Theorem 2.2, given in (9), establishes the iteration complexity of SEG-RR for achieving an error $\mathbb{E}\left[\|z_0^K - z_*\|^2\right] \leq \epsilon$, assuming knowledge of the total number of epochs K. More specifically, after a certain number of epochs K satisfying $K \geq \frac{4\sqrt{120}\lambda^2_{min}(Q)}{L^2_{max}} \log(nK^2)$, the second term dominates in the iteration complexity of SEG-RR (inequality (9)) and thus $\mathbb{E}\left[\|z_0^K - z_*\|^2\right] = \tilde{\mathcal{O}}\left(\frac{1}{nK^2}\right)$. In contrast, after the same number of epochs, the iteration complexity of S-SEG with constant step sizes γ_1 and γ_2 of Hsieh et al. [2020] is equal to $\tilde{\mathcal{O}}\left(\frac{1}{nK}\right)$.

2.4 Monotone Operators

In this part, we focus on the setting where the operator $F(z)$ in the VIP (2) is monotone. In this case, we prove a sublinear convergence of a weighted average $\mathbb{E}\left[\|F(\tilde{z}_0^k)\|^2\right]$ to a neighborhood around the solution z_*. In addition, for step sizes depending on the total number of epochs K, we prove that SEG-RR can reduce the neighborhood and reach any target accuracy $\epsilon > 0$, establishing in this way an iteration complexity of $\mathcal{O}\left(\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right)$ after a certain number of epochs. As a comparison, S-SEG can guarantee convergence to the same specific target accuracy $\epsilon > 0$, only if it is run with large batch sizes.

Theorem 2.3. Suppose that the operator F is monotone and each F_i, $\forall i \in [n]$, is L_i−Lipschitz.

1. Then, the iterates of the SEG-RR algorithm with step sizes $\gamma_2 = 2\gamma_1, \gamma_1 \leq \frac{1}{3\sqrt{2}nL_{max}}$ satisfy:

$$\mathbb{E}\left[\|F(\tilde{z}_0^k)\|^2\right] \leq \frac{\|z_0 - z_*\|^2}{4nG\gamma_1^2 k} + 6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2 \quad (10)$$

2. If SEG-RR is run with step sizes $\gamma_2 = 2\gamma_1$ and $\gamma_1 \leq \min\left\{\frac{1}{3\sqrt{2}nL_{max}}, \frac{1}{(nK)^{\frac{1}{3}}}\right\}$, where K is the total number of epochs the algorithm is run, then:

$$\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right] \leq \frac{9nL_{max}^2\|z_0 - z_*\|^2}{2GK} + \frac{\|z_0 - z_*\|^2}{4Gn^{\frac{1}{3}}K^{\frac{1}{3}}} + \frac{12n^{\frac{1}{3}}L_{max}^2\,(29+n)\,\sigma_*^2}{K^{\frac{2}{3}}} \quad (11)$$

where $\tilde{z}_0^k = \frac{1}{k}\sum_{j=1}^{k} G_j z_0^j$, $G_j = \left(\frac{1}{G}\right)^j$ and $G = 6\left(A + 4L^2 + 1\right)$.

Theorem 2.3 indicates a sublinear convergence for SEG-RR. The convergence is in an average sense, i.e. the weighted average $\tilde{z}_0^k$ of the iterates z_i^k converges to a neighborhood around the solution z_*, which is proportional to the step sizes and the variance σ_*^2 at the optimum. In particular, the neighborhood around the solution z_* decreases as $\mathcal{O}\left(\gamma_1^2 + \gamma_2^2\right)$. Thus, for smaller step sizes γ_1 and γ_2 we expect that the algorithm will converge to a smaller neighborhood around z_*.

In contrast, in the convergence analysis of S-SEG, the neighborhood around the solution cannot be reduced by selecting only the step sizes. More specifically, Gorbunov et al. [2022a] prove the following upper bound (Corollary E.4 for $\gamma_2 = 4\gamma_1$):

$$\mathbb{E}\left[\|F(\bar{z}_0^K)\|^2\right] \leq \frac{\|z_0 - z_*\|^2}{2\gamma_1\gamma_2 K} + 6\sigma_*^2 \quad (12)$$

where $\bar{z}_0^K$ is a different weighted average of the iterates with weights that depend on the stepsize and the parameters of the problem. The second term on the right-hand side of (12) apparently does not depend on the step sizes of the algorithm. Thus, one cannot reduce the neighborhood of convergence around the solution z_* arbitrarily, even by selecting step sizes that depend on the total number of epochs the algorithm is run.

A minibatch of size $\mathcal{O}(K)$ is required according to Gorbunov et al. [2022a] in order for S-SEG to reduce the variance around the optimum and achieve an arbitrary accuracy $\epsilon > 0$. In contrast, SEG-RR can achieve an arbitrary accuracy without the necessity of large batch sizes by selecting step sizes that depend on the total number of epochs, as shown in (11). In particular, after a certain number of epochs $K \geq \mathcal{O}\left(n^{2.5}\right)$ (see equation (92) in Appendix B.3.2), the second term dominates in the right-hand side of inequality (11) and thus SEG-RR achieves an $\mathcal{O}\left(\frac{1}{(nK)^{\frac{1}{3}}}\right)$ accuracy, arbitrarily close to solution. This indicates an intrinsic difference in the batch sizes required in the two methods, S-SEG and SEG-RR, to converge arbitrarily close to the exact solution z_*.

3 Numerical Experiments

In this section, we show the benefits of SEG-RR by performing numerical experiments[2] in strongly monotone quadratic and bilinear minimax problems, as well as on Wasserstein GANs for learning the mean of a multivariate Gaussian distribution.

In particular, we compare SEG-RR, SEG-SO, and IEG with the uniform with-replacement sampling S-SEG (denoted as *SEG* in the plots). For each experiment, we report the average of 5 runs and plot the relative error $\log(\frac{\|z_0^k - z_*\|^2}{\|z_0 - z_*\|^2})$ over the iterations the algorithm is run.

In the strongly monotone setting, similarly to Choudhury et al. [2023], Gorbunov et al. [2022a], Loizou et al. [2021], we consider a quadratic strongly convex strongly concave minimax problem that has the following form:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \frac{x^\top A_i x}{2} + x^\top B_i y - \frac{y^\top C_i y}{2} + a_i^\top x - c_i^\top y$$

[2] The code for reproducing our experimental results is available at https://github.com/emmanouilidisk/Stochastic-ExtraGradient-with-RR.

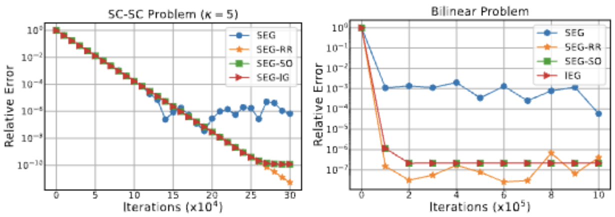

Figure 3: The left plot corresponds to a strongly monotone problem, while the right plot corresponds to a bilinear game. SEG-RR with the theoretical step sizes converges to a smaller relative error compared to the other variants of SEG.

while in the affine regime, we focus on the following two-player bilinear zero-sum game:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}x^\top B_i y + a_i^\top x - c_i^\top y$$

We provide details regarding the way that the matrices A_i, B_i, C_i and the vectors a_i, c_i are sampled in the above problems along with a full description of our experimental setup in Appendix D.

Theoretical step sizes. In the first experiment, we focus on validating Theorems 2.1 and 2.2 by running SEG-RR using the step sizes proposed in our analysis. In Figure 3, we observe that both in the strongly monotone and the bilinear case SEG-RR with constant step sizes converges linearly to a neighborhood around the minimax solution z^*, verifying our theoretical results.

In addition, Figure 3 shows that the three without replacement strategies SEG-RR, SEG-SO, and IEG outperform the uniform with-replacement sampling counterpart of SEG for the same number of epochs/iterations. In our experiments, we also observe that SEG-RR reaches the same neighborhood of convergence (if not smaller) compared to SEG-SO and IEG. We have run experiments, also, for problems with different Lipschitz constants and have observed similar behavior of convergence for SEG-RR. The additional experiments for different Lipschitz parameters can be found in Appendix D.2.

Beyond Theory: Larger step sizes. In the second set of experiments, we investigate the behavior of SEG-RR with larger step sizes than the ones that our theory predicts. That is, we use larger step sizes proposed in previous analyses of S-SEG and compare SEG-RR and S-SEG using these step sizes selection. In particular, for strongly monotone problems, we run experiments for the step sizes proposed in the analysis of S-SEG from Gorbunov et al. [2022a] where $\gamma_1 = \frac{1}{6L_{max}}$ and $\gamma_2 = 4\gamma_1$ while for bilinear games, we use the step sizes $\gamma_1 = \frac{0.1}{(t+19)^{r_\eta}}, \gamma_2 = \frac{1}{(t+19)^{r_\gamma}}$ where $r_\gamma = 0, r_\eta = 0.7$ suggested in the analysis of SEG for bilinear games in Hsieh et al. [2020].

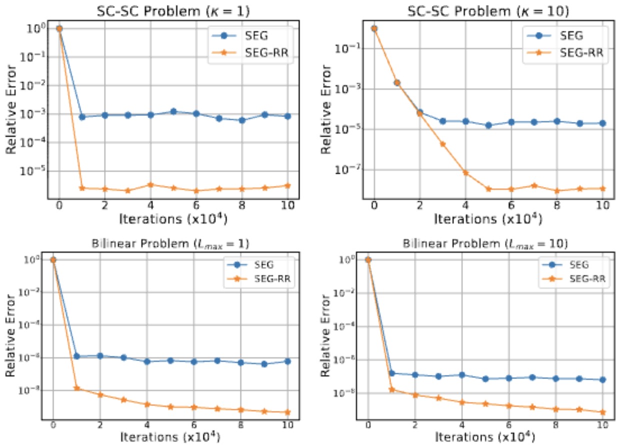

Figure 4: First-row: SC-SC problem. Second-row: Bilinear Game. SEG-RR outperforms SEG in problems with different condition numbers (step size used in SC-SC problem as in Gorbunov et al. [2022a], while step size used in Bilinear Game as in Hsieh et al. [2020]).

In Figure 4, we observe that SEG-RR achieves convergence to a smaller neighborhood than S-SEG for both strongly monotone and bilinear problems. We have, also, conducted additional experiments for more step size and problems with different Lipschitz parameters. We refer the interested reader to Appendix D.2 for a dedicated section.

In all of the experiments, SEG-RR achieves at least as good (if not better) performance than S-SEG, advocating for the use of random reshuffling in practical scenarios, even with step size larger than the ones in our theoretical convergence guarantees.

Wasserstein GANs. In our last experiment, we train a Wasserstein GAN (WGAN) [Arjovsky et al., 2017] for learning the mean of a Multivariate Gaussian distribution. In this scenario, the optimization objective of the WGAN has the following form:

$$\inf_{\theta}\sup_{w}\mathbb{E}_{x\sim N(\mu,\Sigma)}\left[\langle w, x\rangle\right] - \mathbb{E}_{z\sim N(0,\Sigma)}\left[\langle w, z+\theta\rangle\right]$$

In this setting, the discriminator is a linear function $D(x; w) = \langle w, x\rangle$ of the parameter $w \in \mathbb{R}^d$, where the input data point is denoted by x. On the other hand, the generator takes as input a random noise vector $z \sim N(0, \frac{1}{10}I)$ in $\mathbb{R}^d$ and outputs the vector $G(z; \theta) = z + \theta$, which is a linear function of the parameter $\theta \in \mathbb{R}^d$. The goal of the generator is to find the mean μ of the

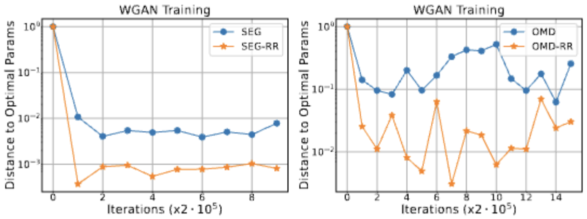

Figure 5: Left: WGAN trained with SEG-RR or S-SEG (denoted as SEG). Right plot: WGAN trained with OMD-RR or OMD. Random reshuffling helps the generator converge closer to the mean $\mu = [3, 4]^T$ of the Gaussian than with-replacement sampling for either the SEG or OMD algorithm.

underlying true distribution $\mathcal{N}(\mu, \Sigma)$, where $\mu = [3, 4]^T$ and $\Sigma = \frac{1}{10} I$.

For the comparison of S-SEG and SEG-RR, we train the WGAN with each one of the two methods. We use the same constant step size for both algorithms with $\gamma_2 = 4\gamma_1, \gamma_1 = 0.01$ being the extrapolation and update step size respectively in both the generator and the discriminator. Figure 5 shows clearly that the generator trained with SEG-RR is able to converge closer to the optimal weights than the generator trained with S-SEG.

Lastly, as in Daskalakis et al. [2018], we train a WGAN with the use of Optimistic Mirror Descent (OMD). Aiming to see the effect of random reshuffling even for this algorithm, we train the WGAN using (i) uniform with-replacement sampling OMD and (ii) OMD with random reshuffling (OMD-RR). We let the step size of the generator and the discriminator be $\gamma_G = 0.02, \gamma_D = 0.01$ respectively. In Figure 5, we observe that random reshuffling allows the OMD algorithm to achieve a smaller distance from the generator's optimal parameters, indicating the benefits of using random reshuffling on top of more popular algorithms.

4 Conclusion

We analyze SEG-RR for strongly monotone, affine, and monotone VIPs. We show that SEG equipped with without-replacement samplings can outperform the iteration complexity of S-SEG after a certain number of epochs. Additionally, in the monotone case, we prove that without-replacement samplings allow the algorithm to converge to an arbitrary accuracy $\epsilon > 0$ without the necessity of having large batch sizes. We aspire that our proof techniques will be a starting point for further results in the field of without-replacement samplings for solving VIPs. In this scope, extending the convergence analysis of SEG-RR to structured non-monotone settings, establishing convergence guarantees for the Stochastic Past ExtraGradient (SPEG) [Choudhury et al., 2023] with random reshuffling, and developing the random reshuffling literature for distributed VIPs [Beznosikov et al., 2022, Zhang et al., 2024] are exciting open research questions that remain to be addressed in the future.

Acknowledgements

Konstantinos Emmanouilidis acknowledges support from a MINDS Fellowship funded by grant NSF 1934979 "HDR TRIPODS: Institute for the Foundations of Graph and Deep Learning." René Vidal acknowledges the support of the NSF-Simons grant 2031985 "Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning." Nicolas Loizou acknowledges support from CISCO Research.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. Sgd with shuffling: optimal rates without component convexity and large epoch requirements. In *NeurIPS*, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. In *NeurIPS*, 2022.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *AISTATS*, 2023.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. 2009.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks*, 2012.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In *NeurIPS*, 2020.

Xufeng Cai, Cheuk Yin Lin, and Jelena Diakonikolas. Empirical risk minimization with shuffled sgd: A primal-dual perspective and improved bounds. *arXiv:2306.12498*, 2023.

George HG Chen and R Tyrrell Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Hanseul Cho and Chulhee Yun. SGDA with shuffling: faster convergence for nonconvex-pł minimax optimization. In *ICLR*, 2023.

Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. In *NeurIPS*, 2023.

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR*, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *AISTATS*, 2022a.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *AISTATS*, 2022b.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *NeurIPS*, 2022c.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *AISTATS*, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. Sgd: General analysis and improved rates. In *AISTATS*, 2019.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186: 49–84, 2021.

Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *ICML*, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS*, 2020.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M Gower, and Peter Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *Journal of Optimization Theory and Applications*, 199(2):499–540, 2023.

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

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020a.

Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *ICML*, 2020b.

Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *ICML*, 2020.

Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. In *NeurIPS*, 2021.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *NeurIPS*, 2020a.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *AISTATS*, 2020b.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NeurIPS*, 2016.

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

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *COLT*, 2020.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *ICLR*, 2023.

Jingkang Wang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, and Bo Li. Adversarial attack generation empowered by min-max optimization. In *NeurIPS*, 2021.

Yaodong Yu, Tianyi Lin, Eric V Mazumdar, and Michael Jordan. Fast distributionally robust learning with variance-reduced min-max optimization. In *AISTATS*, 2022.

Siqi Zhang, Sayantan Choudhury, Sebastian U Stich, and Nicolas Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. In *ICLR*, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes, they are included in the Supplemental Material.]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Stochastic Extragradient with Random Reshuffling: Improved Convergence for Variational Inequalities Supplementary Material

The Supplementary Material is organized as follows: In Section A, we provide some preparatory lemmas and propositions. Section B presents the proofs of our main theorems for SEG-RR. Section C provides further convergence guarantees for SEG-RR with decreasing/switching step size. We, also, explain how the convergence of SEG-SO and IEG is obtained as a corollary of SEG-RR analysis. In Section D, we describe in detail our experimental setup and provide additional experiments.

Contents

A Preparatory Lemmas & Propositions

We start by providing the basic notation and some useful inequalities we use in our proofs, as well as essential preliminaries on variational inequalities. In subsection A.4, a proposition about random reshuffling is provided that is critical in the analysis of stochastic algorithms equipped with without-replacement sampling. In subsection A.5, we state a proposition for bounding the variance of the stochastic oracles F_i, and in subsection A.6, we conclude this section with lemmas that will be necessary in the proofs of our main theorems.

A.1 Notation

We start by introducing the notation that will be useful for stating formally our main results. Let $[n] = \{1, ..., n\}$ and $\mathbb{S}_n$ be the symmetric group of $[n]$. We denote with π^k the permutation of the random reshuffling algorithm at epoch k and with π_i^k the i-th element of the permutation π^k, for $0 \leq i \leq n-1$. The i-th iterate of the algorithm at the k-th epoch will be indicated by z_i^k. The expectation over the uniform distribution of all permutations $\mathcal{D} = \mathcal{U}(\mathbb{S}_n)$ condition on the natural filtration $\mathcal{F}_k$ of z_0^k is denoted by $\mathbb{E}\left[\cdot\middle|\mathcal{F}^k\right] = \mathbb{E}\big[\cdot \mid \mathcal{F}_k\big]$. The expectation taking into account all the stochasticity of the algorithm is denoted by $\mathbb{E}\left[\cdot\right]$.

We, also, denote the extrapolation and update step of the SEG-RR algorithm with

$$\bar{z}_{i-1}^k = z_{i-1}^k - \gamma_2 F_{\pi_{i-1}^k}(z_{i-1}^k) \tag{13}$$

$$z_i^k = z_{i-1}^k - \gamma_1 F_{\pi_{i-1}^k}(\bar{z}_{i-1}^k) \tag{14}$$

as well as an additional variable useful in our proofs with

$$\hat{z}_i^k = z_i^k - \gamma_2 F(z_i^k) \tag{15}$$

A.2 Useful Inequalities

In this section, we provide inequalities that will be useful in our proofs

$$\left\|\sum_{i=1}^n x_i\right\|^2 \leq n\sum_{i=1}^n \|x_i\|^2 \tag{16}$$

$$\|a-b\|^2 \geq \frac{1}{2}\|a\|^2 - \|b\|^2 \tag{17}$$

$$\langle a, b\rangle = \frac{1}{2}\left[\|a\|^2 + \|b\|^2 - \|a-b\|^2\right] \tag{18}$$

$$e^{-x} \geq 1-x, \forall x \geq 0 \tag{19}$$

Using Jensen inequality for $f(x) = \|x\|^2$ yields $\forall t \in [0,1]$ the below inequality:

$$\|a+b\|^2 = \left\|\frac{t}{t}a + \frac{1-t}{1-t}b\right\|^2 \leq t\left\|\frac{a}{t}\right\|^2 + (1-t)\left\|\frac{b}{1-t}\right\|^2 = \frac{1}{t}\|a\|^2 + \frac{1}{1-t}\|b\|^2$$

$$\iff \|a+b\|^2 \leq \frac{1}{t}\|a\|^2 + \frac{1}{1-t}\|b\|^2 \tag{20}$$

Substituting $t = 1 - \frac{1}{2}\gamma_1 n\mu \in [0,1]$ in inequality (20), we have that

$$\|a+b\|^2 \leq \frac{1}{1-\frac{1}{2}\gamma_1 n\mu}\|a\|^2 + \frac{2}{\gamma_1 n\mu}\|b\|^2 \tag{21}$$

Similarly, substituting $t = 1 - \gamma_1 n(\lambda_{min}^+(Q) - \gamma_2 L_{max}^2) \in [0,1]$ in inequality (20) we get

$$\|a+b\|^2 \leq \frac{1}{1-\gamma_1 n(\lambda_{min}^+(Q) - \gamma_2 L_{max}^2)}\|a\|^2 + \frac{1}{\gamma_1 n(\lambda_{min}^+(Q) - \gamma_2 L_{max}^2)}\|b\|^2 \tag{22}$$

A.3 Min-Max Optimization and Variational Inequalities

In the following, we establish the connection between min-max optimization problems and VIPs. We focus on bilinear min-max optimization problems and explain how they can be cast as a special case of affine VIPs. Similar connections can be established for strongly convex-strongly concave and convex-concave min-max optimization problems.

Given a bilinear game of the following form

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x,y) := \frac{1}{n} \sum_{i=0}^{n-1} x^\top B_i y + a_i^\top x - c_i^\top y, \tag{23}$$

letting $Q_i = \begin{bmatrix} 0 & B_i \\ -B_i^T & 0 \end{bmatrix}$, $b_i = \begin{bmatrix} a_i \\ c_i \end{bmatrix}$ one can observe that the problem corresponds to a variational inequality with an affine operator $F(z) = \frac{1}{n} \sum\limits_{i=0}^{n-1} Q_i z + b_i$. Similarly, when the minimax problem is (strongly) convex-(strongly) concave, then the associated variational inequality operator $F(z)$ is (strongly) monotone. Thus, minimax optimization problems are a special case of the problems encapsulated under the more general framework of variational inequalities.

A.4 Proposition about Random Reshuffling

We, next, state a proposition about random reshuffling that will turn out to be helpful in deriving the lemmas of Section A.6.

Proposition A.1 (Mishchenko et al. [2020a]). Let $\{X_1, \ldots, X_n\} \in \mathbb{R}^d$ be a population of n random vectors, $\mu \triangleq \frac{1}{n}\sum_{i=1}^n X_i$ the population average and $\sigma^2 \triangleq \frac{1}{n}\sum_{i=1}^n \|X_i - \mu\|^2$ the population variance.
Take a sample $\{X_{\pi_0}, ..., X_{\pi_{d-1}}\}$ of $d \in [n]$ random vectors from $\{X_1, \ldots, X_n\}$ uniformly at random without replacement and let $\bar{X} = \frac{1}{d}\sum_{i=0}^{d-1} X_{\pi(i)}$ be the sample average and $\mathrm{Var}(\bar{X})$ the sample variance.

Then, we have that:

$$\mathbb{E}_{\pi \in \mathcal{S}}\left[\left\|\bar{X} - \mu\right\|^2\right] = \mathbb{E}_{\pi \in \mathcal{S}}\left[\left\|\frac{1}{d}\sum_{i=0}^{d-1} X_{\pi_i} - \frac{1}{n}\sum_{i=1}^{n} X_i\right\|^2\right] = \frac{n-d}{d(n-1)}\sigma^2. \tag{24}$$

where the expectation is taken with respect to the set $\mathcal{S}$, which is the set of permutations of length d of $[n]$.

Proof. We first establish the identity $\mathrm{cov}(X_{\pi_i}, X_{\pi_j}) = -\frac{\sigma^2}{n-1}, \forall i \neq j$ as follows:

$$\begin{aligned}
\mathrm{cov}(X_{\pi_i}, X_{\pi_j}) &= \frac{1}{n(n-1)} \sum_{l=1}^{n} \sum_{m=1, m\neq l}^{n} \mathbb{E}\left[X_l - \mu, X_m - \mu\right] \\
&= \frac{1}{n(n-1)} \sum_{l=1}^{n} \sum_{m=1}^{n} \mathbb{E}\left[X_l - \mu, X_m - \mu\right] - \frac{1}{n(n-1)} \sum_{l=1}^{n} \|X_l - \mu\|^2 \\
&= \frac{1}{n(n-1)} \sum_{l=1}^{n} \mathbb{E}\left[X_l - \mu, \sum_{m=1}^{n}(X_m - \mu)\right] - \frac{\sigma^2}{n-1} = -\frac{\sigma^2}{n-1}.
\end{aligned}$$

We, now, turn to the formula for sample variance:

$$\begin{aligned}
\mathrm{Var}(\bar{X}) = \mathbb{E}_{\pi \in \mathcal{S}}\left[\left\|\bar{X} - \mu\right\|^2\right] &= \frac{1}{d^2} \sum_{i=0}^{d-1} \sum_{j=0}^{d-1} \mathrm{cov}(X_{\pi_i}, X_{\pi_j}) \\
&= \frac{1}{d^2}\left[\sum_{i=0}^{d-1} \sum_{j=i}^{d-1} \mathrm{cov}(X_{\pi_i}, X_{\pi_j}) + \sum_{i=0}^{d-1} \sum_{j=0, j\neq i}^{d-1} \mathrm{cov}(X_{\pi_i}, X_{\pi_j})\right]
\end{aligned}$$

$$= \frac{1}{d^2}\left[\sum_{i=0}^{d-1}\sum_{j=i}^{d-1}\mathrm{Var}(X_{\pi_i}) + \sum_{i=0}^{d-1}\sum_{j=0, j\neq i}^{d-1}\mathrm{cov}(X_{\pi_i}, X_{\pi_j})\right]$$

We, thus, continue our arithmetic manipulations

$$\mathbb{E}_{\pi\in\mathcal{S}}\left[\left\|\bar{X}-\mu\right\|^2\right] = \tfrac{1}{d^2}\left(d\cdot\sigma^2 - d(d-1)\tfrac{\sigma^2}{n-1}\right) = \tfrac{n-d}{d(n-1)}\sigma^2 \tag{25}$$

to conclude with the promised equation in the statement of this proposition. □

A.5 Variance of Stochastic Oracles

We provide a proposition for bounding the variance of the stochastic oracles F_i. As mentioned in the main paper, our approach follows a recent line of work [Choudhury et al., 2023, Gorbunov et al., 2022a, Gower et al., 2021, 2019, Khaled et al., 2023, Loizou et al., 2020, 2021] that uses the Lipschitz assumption to provide closed-form expressions for the upper bound on the variance.

Proposition A.2. If each F_i is L_i-Lipschitz, then $\forall z\in\mathbb{R}^d$ the following holds

$$\frac{1}{n}\sum_{i=1}^{n}\|F_i(z)-F(z)\|^2 \leq A\|z-z_*\|^2 + 2\sigma_*^2$$

where $A = \frac{2}{n}\sum\limits_{i=1}^{n} L_i^2$ and $\sigma_*^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\|F_i(z_*)\|^2$.

Proof.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|F_i(z)-F(z)\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\|F_i(z)\|^2 - \frac{2}{n}\sum_{i=1}^{n}\langle F_i(z), F(z)\rangle + \frac{1}{n}\sum_{i=1}^{n}\|F(z)\|^2\\
&= \frac{1}{n}\sum_{i=1}^{n}\|F_i(z)\|^2 - 2\|F(z)\|^2 + \|F(z)\|^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|F_i(z)\|^2\\
&\overset{(16)}{\leq} \frac{2}{n}\sum_{i=1}^{n}\|F_i(z)-F_i(z_*)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|F_i(z_*)\|^2\\
&= A\|z-z_*\|^2 + 2\sigma_*^2
\end{aligned}$$

where $A = \frac{2}{n}\sum_{i=1}^{n} L_i^2$ and $\sigma_*^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\|F_i(z_*)\|^2$. □

A.6 Useful Lemmas

In this section, we provide some necessary preparatory lemmas that will be crucial for proving our main results. We start with a lemma that bounds the distance between the operator F and $\frac{1}{d}\sum_{j=0}^{d-1}F_{\pi_j}$ with $d\in[n]$.

Lemma A.3. Suppose that each $F_i, i\in[n]$ is L_i-Lipschitz. For any fixed $d\in[n]$ and $z\in\mathbb{R}^d$, the following inequality holds:

$$d^2\mathbb{E}_{\pi\in\mathcal{S}}\left[\left\|\frac{1}{d}\sum_{j=0}^{d-1}F_{\pi_j}(z)-F(z)\right\|^2\right] \leq \frac{d(n-d)}{n-1}\left(A\|z-z_*\|^2+2\sigma_*^2\right) \tag{26}$$

where $\mathcal{S}$ is the set of all permutations of the set $[n]$ of length d, $A = \frac{2}{n}\sum_{i=1}^{n} L_i^2$ and $\sigma_*^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\|F_i(z_*)\|^2$.

Proof. First, we substitute in Proposition A.1 $X_i \leftarrow F_i(z_0^k), i \in [n]$ and fix an integer $d \in [n]$. Next, we draw a permutation with d elements uniformly at random from the set of all permutations of $[n]$ with d elements, i.e. $\pi \sim \mathcal{U}(\mathcal{S})$. In other words, let $X_{\pi_0}, \ldots X_{\pi_{d-1}}$ be sampled uniformly without replacement from $\{X_1, \ldots, X_n\}$. Then, we have that the quantities $\bar{X}, \mu$ from Proposition A.1 are equal to:

$$\bar{X} = \frac{1}{d}\sum_{j=0}^{d-1} F_{\pi_j}(z), \qquad \mu = \frac{1}{n}\sum_{j=1}^{n} F_j(z) = F(z) \tag{27}$$

From Proposition A.1, thus, we get that:

$$\mathbb{E}_{\pi\in\mathcal{S}}\left[\left\|\bar{X} - \mu\right\|^2\right] = \mathbb{E}_{\pi\in\mathcal{S}}\left[\left\|\frac{1}{d}\sum_{j=0}^{d-1} F_{\pi_j}(z) - F(z)\right\|^2\right] = \frac{n-d}{d(n-1)}\frac{1}{n}\sum_{j=1}^{n}\|F_j(z) - F(z)\|^2$$

Using Proposition A.2, we next bound the sum on the right hand-sight (RHS) as follows:

$$d^2\mathbb{E}_{\pi\in\mathcal{S}}\left[\left\|\frac{1}{d}\sum_{j=0}^{d-1} F_{\pi_j}(z) - F(z)\right\|^2\right] \quad \leq \quad \frac{d(n-d)}{n-1}\left(A\|z - z_*\|^2 + 2\sigma_*^2\right)$$

□

We, next, provide a lemma bounding the average distance of the iterates inside the $k-$th epoch from the initial point z_0^k in the epoch.

Lemma A.4. Assume that each $F_i, i \in [n]$, is L_i-Lipchitz and the step size of SEG-RR satisfy $\gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}$, $\gamma_2 \leq \frac{1}{\sqrt{n(n-1)}L_{max}}$. The iterates of the SEG-RR algorithm satisfy the following bound

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\left\|z_j^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad \left[10n^2L^2 + A(25+n)\right]\gamma_1^2\left\|z_0^k - z_*\right\|^2 + 2(n+25)\gamma_1^2\sigma_*^2 \tag{28}$$

Proof. Using the update rule of SEG-RR in (14), we have that:

$$z_i^k = z_{i-1}^k - \gamma_1 F_{\pi_{i-1}^k}(\bar{z}_{i-1}^k) = z_0^k - \gamma_1\sum_{j=0}^{i-1}(F_{\pi_j^k}(\bar{z}_j^k) - F_{\pi_j^k}(\hat{z}_0^k))$$

We, thus, have that:

$$\begin{aligned}
\left\|z_i^k - z_0^k\right\|^2 &= \gamma_1^2 i^2\left\|\frac{1}{i}\sum_{j=0}^{i-1} F_{\pi_j^k}(\bar{z}_j^k)\right\|^2 \\
&= \gamma_1^2 i^2\left\|\frac{1}{i}\sum_{j=0}^{i-1}\left[F_{\pi_j^k}(\bar{z}_j^k) - F_{\pi_j^k}(z_0^k) + F_{\pi_j^k}(z_0^k)\right] - F(z_0^k) + F(z_0^k)\right\|^2 \\
&\overset{(16)}{\leq} 3\gamma_1^2 i\sum_{j=0}^{i-1}\left\|F_{\pi_j^k}(\bar{z}_j^k) - F_{\pi_j^k}(z_0^k)\right\|^2 + 3\gamma_1^2 i^2\left\|\frac{1}{i}\sum_{j=0}^{i-1}F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 + 3\gamma_1^2 i^2\left\|F(z_0^k)\right\|^2
\end{aligned}$$

Using the Lipschitz property of $F_{\pi_j^k}(z)$, we continue as follows

$$\left\|z_i^k - z_0^k\right\|^2 \quad \leq \quad 3\gamma_1^2 L_{max}^2 i\sum_{j=0}^{i-1}\left\|\bar{z}_j^k - z_0^k\right\|^2 + 3\gamma_1^2 i^2\left\|\frac{1}{i}\sum_{j=0}^{i-1}F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 + 3\gamma_1^2 i^2\left\|F(z_0^k)\right\|^2$$

Substituting the update rule (13) of the extrapolation step of SEG-RR, we get:

$$\begin{aligned}
\left\|z_i^k - z_0^k\right\|^2 &\overset{(13)}{\leq} 3\gamma_1^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| z_j^k - \gamma_2 F_{\pi_j^k}(z_j^k) - z_0^k \right\|^2 + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2 \\
&\overset{(16)}{\leq} 6\gamma_1^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| z_j^k - z_0^k \right\|^2 + 6\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_j^k) \right\|^2 \\
&\quad + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2
\end{aligned}$$

Continuing with further algebraic manipulations, we obtain

$$\begin{aligned}
\left\|z_i^k - z_0^k\right\|^2 &\leq 6\gamma_1^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| z_j^k - z_0^k \right\|^2 + 6\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_j^k) - F_{\pi_j^k}(z_0^k) + F_{\pi_j^k}(z_0^k) \right\|^2 \\
&\quad + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2 \\
&\overset{(16)}{\leq} 6\gamma_1^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| z_j^k - z_0^k \right\|^2 + 12\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_j^k) - F_{\pi_j^k}(z_0^k) \right\|^2 \\
&\quad + 12\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2
\end{aligned}$$

Using the Lipschitz property of $F_{\pi_j^k}(z)$ results to

$$\begin{aligned}
\left\|z_i^k - z_0^k\right\|^2 &\leq 6\gamma_1^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| z_j^k - z_0^k \right\|^2 + 12\gamma_1^2 \gamma_2^2 L_{max}^4 i \sum_{j=0}^{i-1} \left\| z_j^k - z_0^k \right\|^2 \\
&\quad + 12\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2 \\
&\leq 6\gamma_1^2 L_{max}^2 i \sum_{j=0}^{n-1} \left\| z_j^k - z_0^k \right\|^2 + 12\gamma_1^2 \gamma_2^2 L_{max}^4 i \sum_{j=0}^{n-1} \left\| z_j^k - z_0^k \right\|^2 \\
&\quad + 12\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2
\end{aligned}$$

Letting $G_k = \frac{1}{n} \sum\limits_{j=0}^{n-1} \left\| z_j^k - z_0^k \right\|^2$ for brevity, we have that:

$$\begin{aligned}
\left\|z_i^k - z_0^k\right\|^2 &\leq 6\gamma_1^2 L_{max}^2 i (1 + 2\gamma_2^2 L_{max}^2) n G_k + 12\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{i-1} \left\| F_{\pi_j^k}(z_0^k) - F(z_0^k) + F(z_0^k) \right\|^2 \\
&\quad + 3\gamma_1^2 i^2 \left\| \frac{1}{i} \sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2 + 3\gamma_1^2 i^2 \left\| F(z_0^k) \right\|^2 \\
&\overset{(16)}{\leq} 6\gamma_1^2 L_{max}^2 i (1 + 2\gamma_2^2 L_{max}^2) n G_k + 24\gamma_1^2 \gamma_2^2 L_{max}^2 i \sum_{j=0}^{n-1} \left\| F_{\pi_j^k}(z_0^k) - F(z_0^k) \right\|^2
\end{aligned}$$

$$+3\gamma_1^2 i^2 \left\|\frac{1}{i}\sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 + 3\gamma_1^2(i^2+8\gamma_2^2 i^2 L_{max}^2)\left\|F(z_0^k)\right\|^2$$

Taking expectation condition on the filtration $\mathcal{F}_k$ we get:

$$\begin{aligned}\mathbb{E}\left[\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] &\leq 6\gamma_1^2 L_{max}^2 i(1+2\gamma_2^2 L_{max}^2) n G_k + 3\gamma_1^2(i^2+8\gamma_2^2 i^2 L_{max}^2)\left\|F(z_0^k)\right\|^2 \\ &\quad +24\gamma_1^2\gamma_2^2 L_{max}^2 i \mathbb{E}\left[\sum_{j=0}^{i-1}\left\|F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \\ &\quad +3\gamma_1^2 i^2 \mathbb{E}\left[\left\|\frac{1}{i}\sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]\end{aligned} \tag{29}$$

We next bound the last two terms of (29). Using Proposition (A.2), we have that:

$$\begin{aligned}&\frac{1}{n}\sum_{j=0}^{n-1}\left\|F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \leq A\left\|z_0^k - z_*\right\|^2 + 2\sigma_*^2 \\ \iff &\sum_{j=0}^{n-1}\left\|F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \leq An\left\|z_0^k - z_*\right\|^2 + 2n\sigma_*^2\end{aligned}$$

Taking conditional expectation on both sides of the inequality, results to

$$\mathbb{E}\left[\sum_{j=0}^{n-1}\left\|F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \leq An\left\|z_0^k - z_*\right\|^2 + 2n\sigma_*^2 \tag{30}$$

Substituting inequality (30) in (29), we get:

$$\begin{aligned}\mathbb{E}\left[\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] &\overset{(30)}{\leq} 6\gamma_1^2 L_{max}^2 i(1+2\gamma_2^2 L_{max}^2) n G_k + 3\gamma_1^2(i^2+8\gamma_2^2 i^2 L_{max}^2)\left\|F(z_0^k)\right\|^2 \\ &\quad +24\gamma_1^2\gamma_2^2 L_{max}^2 i\left(An\left\|z_0^k - z_*\right\|^2 + 2n\sigma_*^2\right) \\ &\quad +3\gamma_1^2 i^2 \mathbb{E}\left[\left\|\frac{1}{i}\sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \\ &= 6\gamma_1^2 L_{max}^2 i(1+2\gamma_2^2 L_{max}^2) n G_k + 3\gamma_1^2(i^2+8\gamma_2^2 i^2 L_{max}^2)\left\|F(z_0^k)\right\|^2 \\ &\quad +24\gamma_1^2\gamma_2^2 L_{max}^2 Ani\left\|z_0^k - z_*\right\|^2 + 48\gamma_1^2\gamma_2^2 L_{max}^2 ni\sigma_*^2 \\ &\quad +3\gamma_1^2 i^2 \mathbb{E}\left[\left\|\frac{1}{i}\sum_{j=0}^{i-1} F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]\end{aligned} \tag{31}$$

Using Lemma A.3 with $d \leftarrow i$, we can bound the last term in (31) and get:

$$\begin{aligned}\mathbb{E}\left[\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] &\overset{(A.3)}{\leq} 6\gamma_1^2 L_{max}^2 i(1+2\gamma_2^2 L_{max}^2) n \mathbb{E}\left[G_k \middle| \mathcal{F}^k\right] + 3\gamma_1^2\left(1+8\gamma_2^2 L_{max}^2\right) i^2\left\|F(z_0^k)\right\|^2 \\ &\quad +3\gamma_1^2 A\left[\frac{i(n-i)}{n-1} + 8\gamma_2^2 ni L_{max}^2\right]\left\|z_0^k - z_*\right\|^2 \\ &\quad +6\gamma_1^2\sigma_*^2\left[\frac{i(n-i)}{n-1} + 8\gamma_2^2 ni L_{max}^2\right]\end{aligned} \tag{32}$$

Summing over $0 \leq i \leq n-1$ and multiplying with $\frac{1}{n}$, we have that:

$$\begin{aligned}
\mathbb{E}\left[G_k \middle| \mathcal{F}^k\right] = \frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad 3\gamma_1^2 L_{max}^2(1+2\gamma_2^2 L_{max}^2)n(n-1)\mathbb{E}\left[G_k \middle| \mathcal{F}^k\right] + \gamma_1^2 D \left\|F(z_0^k)\right\|^2 \\
&\quad + \gamma_1^2 A\left[\frac{n+1}{2} + 12\gamma_2^2 L_{max}^2 n(n-1)\right]\left\|z_0^k - z_*\right\|^2 \\
&\quad + \gamma_1^2\sigma_*^2\left[(n+1) + 24\gamma_2^2 L_{max}^2 n(n-1)\right] \qquad (33)
\end{aligned}$$

where we used the facts

$$\frac{1}{n}\sum_{i=0}^{n-1} i = \frac{n-1}{2}, \quad \frac{1}{n}\sum_{i=0}^{n-1} i^2 = \frac{(n-1)(2n-1)}{6}, \quad \frac{1}{n}\sum_{i=0}^{n-1}\frac{i(n-i)}{n-1} = \frac{n+1}{6}$$

and let, also, $D = \left[\frac{(1+8\gamma_2^2 L_{max}^2)(n-1)(2n-1)}{2}\right]$ for brevity.

Rearranging the terms in inequality (33), we obtain:

$$\begin{aligned}
[1 - 3n(n-1)\gamma_1^2(1+2\gamma_2^2 L_{max}^2)L_{max}^2]\mathbb{E}\left[G_k\right] \;&\leq\; \gamma_1^2 D\left\|F(z_0^k)\right\|^2 + \gamma_1^2 A\left[\frac{n+1}{2} + 12\gamma_2^2 L_{max}^2 n(n-1)\right]\left\|z_0^k - z_*\right\|^2 \\
&\quad + \gamma_1^2\sigma_*^2\left[(n+1) + 24\gamma_2^2 L_{max}^2 n(n-1)\right]
\end{aligned}$$

Letting $\gamma_2 \leq \frac{1}{L_{max}}$, we get:

$$\begin{aligned}
[1 - 9n(n-1)\gamma_1^2 L_{max}^2]\mathbb{E}\left[G_k\right] \quad &\leq \quad \gamma_1^2 D\left\|F(z_0^k)\right\|^2 \\
&\quad + \gamma_1^2 A\left[\frac{n+1}{2} + 12\gamma_2^2 L_{max}^2 n(n-1)\right]\left\|z_0^k - z_*\right\|^2 \\
&\quad + \gamma_1^2\sigma_*^2\left[(n+1) + 24\gamma_2^2 L_{max}^2 n(n-1)\right]
\end{aligned}$$

By letting $D_1 = [1 - 9n(n-1)\gamma_1^2 L_{max}^2]$ and selecting the stepsize $\gamma_1 < \frac{1}{3\sqrt{n(n-1)}L_{max}}$, we have that $D_1 > 0$ and thus we get that:

$$\begin{aligned}
\mathbb{E}\left[G_k \middle| \mathcal{F}^k\right] \quad &\leq \quad \frac{D\gamma_1^2}{D_1}\left\|F(z_0^k)\right\|^2 \\
&\quad + \frac{A\gamma_1^2}{D_1}\left[\frac{n+1}{2} + 12\gamma_2^2 L_{max}^2 n(n-1)\right]\left\|z_0^k - z_*\right\|^2 \\
&\quad + \frac{\gamma_1^2}{D_1}\left[(n+1) + 24\gamma_2^2 L_{max}^2 n(n-1)\right]\sigma_*^2 \qquad (34)
\end{aligned}$$

Lastly, substituting the definition of G_k we get that:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\left\|z_j^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \frac{D\gamma_1^2}{D_1}\left\|F(z_0^k)\right\|^2 + \frac{A\gamma_1^2}{D_1}\left[\frac{n+1}{2} + 12\gamma_2^2 L_{max}^2 n(n-1)\right]\left\|z_0^k - z_*\right\|^2 \\
&\quad + \frac{\gamma_1^2}{D_1}\left[(n+1) + 24\gamma_2^2 L_{max}^2 n(n-1)\right]\sigma_*^2 \qquad (35)
\end{aligned}$$

Selecting $\gamma_1 \leq \frac{1}{3L_{max}\sqrt{2n(n-1)}}$ we have that:

$$D_1 = 1 - 9n(n-1)L_{max}^2\gamma_1^2 \geq \frac{1}{2} \iff \frac{1}{2D_1} \leq 1 \qquad (36)$$

We, also, have that for $\gamma_2 \leq \frac{1}{\sqrt{n(n-1)}L_{max}}$ we can upper bound D as follows

$$D = \frac{(1+8\gamma_2^2 L_{max}^2)(n-1)(2n-1)}{2} \leq \frac{5(n-1)(2n-1)}{2} \leq 5n^2 \qquad (37)$$

Substituting the bounds (36), (37) to (35), we obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\left\|z_j^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] \quad &\overset{(36),(37)}{\leq} \quad 10n^2\gamma_1^2\left\|F(z_0^k)\right\|^2 + 2A\gamma_1^2\left[\frac{n+1}{2} + 12\gamma_2^2L_{max}^2 n(n-1)\right]\left\|z_0^k - z_*\right\|^2 \\
&\qquad\quad +2\gamma_1^2\left[(n+1) + 24\gamma_2^2L_{max}^2 n(n-1)\right]\sigma_*^2 \\
&\overset{\gamma_2\leq\frac{1}{\sqrt{n(n-1)}L_{max}}}{\leq} \quad 10n^2\gamma_1^2\left\|F(z_0^k)\right\|^2 + A\gamma_1^2(25+n)\left\|z_0^k - z_*\right\|^2 + 2\gamma_1^2(n+25)\sigma_*^2 \\
&\overset{(3)}{\leq} \quad \left[10n^2L^2 + A(25+n)\right]\gamma_1^2\left\|z_0^k - z_*\right\|^2 + 2(n+25)\gamma_1^2\sigma_*^2
\end{aligned}$$

□

In the next Lemma, we bound a term that appears in the proofs of our theorems.

Lemma A.5. Assume each $F_i, i \in [n]$, is L_i-Lipschitz. If the extrapolation stepsize of SEG-RR satisfies $\gamma_2 \leq \frac{1}{L_{max}}$, then the following bound holds:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad & 6L_{max}^2\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] \\
& +3L_{max}^2\gamma_2^2\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]
\end{aligned}$$

Proof. Using the Lipschitz property of $F_i, \forall i \in [n], k \geq 0$, we have that

$$\begin{aligned}
\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right\|^2 \quad &\leq \quad L_{max}^2\sum_{i=0}^{n-1}\left\|\bar{z}_i^k - \hat{z}_0^k\right\|^2 \\
&\overset{(13)}{=} \quad L_{max}^2\sum_{i=0}^{n-1}\left\|z_i^k - \gamma_2F_{\pi_i^k}(z_i^k) - z_0^k + \gamma_2F(z_0^k) + \gamma_2F_i(z_0^k) - \gamma_2F_i(z_0^k)\right\|^2 \\
&\overset{(16)}{\leq} \quad 3L_{max}^2\sum_{i=0}^{n-1}\left\|z_i^k - z_0^k\right\|^2 + 3L_{max}^2\gamma_2^2\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_i^k) - F_{\pi_i^k}(z_0^k)\right\|^2 \\
&\qquad\quad +3L_{max}^2\gamma_2^2\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2 \\
&\overset{(3)}{\leq} \quad 3L_{max}^2(1+\gamma_2^2L_{max}^2)\sum_{i=0}^{n-1}\left\|z_i^k - z_0^k\right\|^2 + 3L_{max}^2\gamma_2^2\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2
\end{aligned}$$

Taking expectation condition on the filtration $\mathcal{F}_k$ and using Proposition A.2, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad & 3L_{max}^2(1+\gamma_2^2L_{max}^2)\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] \\
& +3L_{max}^2\gamma_2^2\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]
\end{aligned}$$

For $\gamma_2 \leq \frac{1}{L_{max}}$, we get:

$$\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \leq 6L_{max}^2\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right] + 3L_{max}^2\gamma_2^2\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]$$

□

B Proofs for SEG-RR

B.1 Proofs for Strongly Monotone Case

B.1.1 Lemma for Iterates in Strongly Monotone Case

Lemma B.1. For SEG-RR if F is $\mu-$strongly monotone and Assumption 3 holds, we have the following bound:

$$\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 \quad \leq \quad \left(1-\frac{1}{2}\gamma_1 n\mu\right)^2 \|z_0^k - z_*\|^2 + U\|z_0^k - z_*\|^2$$

where $\hat{z}_0^k = z_0^k - \gamma_2 F(z_0^k)$ and $U = \Big\{\gamma_1^2 n^2(2L^2 - \frac{\mu^2}{4}) + 2\gamma_1\gamma_2 n L^2\left[-1+\gamma_2(\gamma_1 n L^2 + \gamma_2 L^2 + \mu)\right]\Big\}$.

Proof. We have that:

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &= \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle z_0^k - z_*, F(\hat{z}_0^k)\rangle + \gamma_1^2 n^2\|F(\hat{z}_0^k)\|^2 \\
&= \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle z_0^k - \gamma_2 F(z_0^k) - z_*, F(\hat{z}_0^k)\rangle \\
&\quad +\gamma_1^2 n^2\|F(\hat{z}_0^k)\|^2 - 2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(15)}{=} \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle + \gamma_1^2 n^2\|F(\hat{z}_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(1.2)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|\hat{z}_0^k - z_*\|^2 + \gamma_1^2 n^2\|F(\hat{z}_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&= \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|\hat{z}_0^k - z_*\|^2 + \gamma_1^2 n^2\|F(\hat{z}_0^k) - F(z_0^k) + F(z_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(16)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|\hat{z}_0^k - z_*\|^2 + 2\gamma_1^2 n^2\|F(\hat{z}_0^k) - F(z_0^k)\|^2 \\
&\quad +2\gamma_1^2 n^2\|F(z_0^k)\|^2 - 2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle
\end{aligned}$$

Using the Lipschitz property of F, we get

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &\overset{(3)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|\hat{z}_0^k - z_*\|^2 + 2\gamma_1^2 n^2 L^2\|\hat{z}_0^k - z_0^k\|^2 \\
&\quad +2\gamma_1^2 n^2\|F(z_0^k)\|^2 - 2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(18)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|\hat{z}_0^k - z_*\|^2 + 2\gamma_1^2 n^2 L^2\|\hat{z}_0^k - z_0^k\|^2 \\
&\quad +2\gamma_1 n(\gamma_1 n - \gamma_2)\|F(z_0^k)\|^2 + 2\gamma_1\gamma_2 n\|F(z_0^k) - F(\hat{z}_0^k)\|^2 \\
&\overset{(3)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|\hat{z}_0^k - z_*\|^2 + 2\gamma_1 n L^2(\gamma_1 n + \gamma_2)\|\hat{z}_0^k - z_0^k\|^2 \\
&\quad +2\gamma_1 n(\gamma_1 n - \gamma_2)\|F(z_0^k)\|^2
\end{aligned}$$

Substituting the definition of $\hat{z}_0^k$ we have

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &\overset{(15)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\mu\|z_0^k - \gamma_2 F(z_0^k) - z_*\|^2 \\
&\quad +2\gamma_1 n(\gamma_1 n - \gamma_2 + \gamma_2^2 L^2(\gamma_1 n + \gamma_2))\|F(z_0^k)\|^2
\end{aligned}$$

We, next, make use of inequality (17) to obtain

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &\overset{(17)}{\leq} (1-\gamma_1 n\mu)\|z_0^k - z_*\|^2 \\
&\quad +2\gamma_1 n\left[\gamma_1 n - \gamma_2 + \gamma_2^2 L^2(\gamma_1 n + \gamma_2) + \gamma_2^2\mu\right]\|F(z_0^k)\|^2 \\
&= \left(1-\frac{1}{2}\gamma_1 n\mu\right)^2 \|z_0^k - z_*\|^2 - \frac{1}{4}\gamma_1^2 n^2\mu^2\|z_0^k - z_*\|^2 \\
&\quad +2\gamma_1 n\left[\gamma_1 n - \gamma_2 + \gamma_2^2 L^2(\gamma_1 n + \gamma_2) + \gamma_2^2\mu\right]\|F(z_0^k)\|^2
\end{aligned}$$

$$\overset{(3)}{\leq} \quad \left(1-\frac{1}{2}\gamma_1 n\mu\right)^2 \|z_0^k - z_*\|^2 + U\|z_0^k - z_*\|^2$$

where $U = \left\{\gamma_1^2 n^2(2L^2 - \frac{\mu^2}{4}) + 2\gamma_1\gamma_2 nL^2\left[-1+\gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right]\right\}$. □

Lemma B.2. If the step size of the SEG-RR algorithm satisfy $\gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$, $\gamma_2 = 2\gamma_1$ then the following holds:

$$\frac{U}{1-\frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu} \quad \leq \quad \frac{\gamma_1 n\mu}{4} \tag{38}$$

where the constants are $C = 2\left[(25+n)A + 10n^2L^2\right]\gamma_1^2 + A\gamma_2^2$,
$U = \left\{\gamma_1^2 n^2(2L^2 - \frac{\mu^2}{4}) + 2\gamma_1\gamma_2 nL^2\left[-1+\gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right]\right\}$.

Proof. We have that:

$$\begin{aligned}
\frac{U}{1-\frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu} &\leq \frac{\gamma_1 n\mu}{4} \\
\iff \frac{\gamma_1^2 n^2(2L^2 - \frac{\mu^2}{4}) + 2\gamma_1\gamma_2 nL^2\left[-1+\gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right]}{1-\frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu} &\leq \frac{\gamma_1 n\mu}{4} \\
\iff \frac{\gamma_1 n(2L^2 - \frac{\mu^2}{4}) + 2\gamma_2 L^2\left[-1+\gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right]}{1-\frac{\gamma_1 n\mu}{2}} + \frac{6CL_{max}^2}{\mu} &\leq \frac{\mu}{4}
\end{aligned}$$

Rearranging the terms we get

$$\begin{aligned}
& \gamma_1 n\left(2L^2 - \frac{\mu^2}{8}\right) + 2\gamma_2 L^2\left[-1+\gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right] + \frac{6CL_{max}^2}{\mu}\left(1-\frac{\gamma_1 n\mu}{2}\right) - \frac{\mu}{4} \leq 0 \\
\iff \quad & \gamma_1 n\left(2L^2 - \frac{\mu^2}{8}\right) + 2\gamma_2 L^2\left[-1+\gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right] \\
& + \frac{6\left\{2\left[(25+n)A + 10n^2L^2\right]\gamma_1^2 + A\gamma_2^2\right\}L_{max}^2}{\mu}\left(1-\frac{\gamma_1 n\mu}{2}\right) - \frac{\mu}{4} \leq 0 \\
\overset{\gamma_2=2\gamma_1}{\iff} \quad & \gamma_1 n\left(2L^2 - \frac{\mu^2}{8}\right) + 4L^2\gamma_1\left\{-1+2\gamma_1[\gamma_1(n+2)L^2 + \mu]\right\} \\
& + \frac{6\left[(54+2n)A + 20n^2L^2\right]\gamma_1^2 L_{max}^2}{\mu}\left(1-\frac{\gamma_1 n\mu}{2}\right) - \frac{\mu}{4} \leq 0
\end{aligned} \tag{39}$$

For $\gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}$, we have that

$$\begin{aligned}
\left\{-1+2\gamma_1[\gamma_1(n+2)L^2 + \mu]\right\} \quad & \overset{\gamma_1\leq\frac{1}{3\sqrt{2}nL_{max}}}{\leq} \quad \left\{-1 + \frac{2(n+2)L^2}{18n(n-1)L_{max}^2} + \frac{\mu}{3\sqrt{2n(n-1)}L_{max}}\right\} \\
& \leq \quad \left[-1 + \frac{1}{9(n-1)L_{max}^3\sqrt{2n(n-1)}} + \frac{\mu}{9n(n-1)L_{max}^2}\right] \\
& \leq \quad -1 + \frac{1}{9} + \frac{1}{9} = -\frac{7}{9}
\end{aligned} \tag{40}$$

Thus, using (40) and the fact that $(1-\frac{\gamma_1 n\mu}{2}) \leq 1$ in (39) it suffices to ensure that

$$\begin{aligned}
\gamma_1 n\left(2L^2 - \frac{\mu^2}{8}\right) - \frac{28L^2}{9}\gamma_1 + \frac{6\left[(54+2n)A + 20n^2L^2\right]L_{max}^2}{\mu}\gamma_1^2 - \frac{\mu}{4} \leq 0 \\
\gamma_1\left[2nL^2 - \frac{28L^2}{9} - \frac{n\mu^2}{8}\right] + \frac{6\left[(54+2n)A + 20n^2L^2\right]L_{max}^2}{\mu}\gamma_1^2 - \frac{\mu}{4} \leq 0
\end{aligned}$$

Thus, it suffices to ensure that:

$$2nL^2\gamma_1 + \frac{6\left[(54+2n)A + 20n^2L^2\right]L_{max}^2}{\mu}\gamma_1^2 - \frac{\mu}{4} \leq 0 \tag{41}$$

In order, now, to derive a simple expression for the stepsize γ_1, instead of solving the quadratic inequality (41), we choose γ_1 such that

$$\begin{array}{ccc} 2nL^2\gamma_1 - \frac{\mu}{8} \leq 0 & \text{and} & +\frac{6\left[(54+2n)A + 20n^2L^2\right]L_{max}^2}{\mu}\gamma_1^2 - \frac{\mu}{8} \leq 0 \\ \gamma_1 \leq \frac{\mu}{16nL^2} & \text{and} & \gamma_1 \leq \frac{\mu}{\sqrt{48\left[(54+2n)A + 20n^2L^2\right]L_{max}^2}} \leq 0 \end{array} \tag{42}$$

Using the fact that $A = \frac{2}{n}\sum\limits_{i=0}^{n-1} L_i^2 \leq 2L_{max}^2$ and $L^2 \leq L_{max}^2$, we observe that it suffices

$$\begin{array}{ccc} \gamma_1 \leq \frac{\mu}{16nL^2} & \text{and} & \gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}} \leq 0 \\ & \Longleftrightarrow & \gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}} \end{array}$$

Lastly, incorporating the initial constraint that $\gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}$, it suffices to choose $\gamma_2 = 2\gamma_1$ and

$$\gamma_1 \quad = \quad \min\left\{\frac{1}{3\sqrt{2n(n-1)}L_{max}}, \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}\right\} = \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$$

□

B.1.2 Proof of Theorem 2.1

Proof. Denote with $\bar{z}_i^k = z_i^k - \gamma_2 F_{\pi_i^k}(z_i^k)$ the extrapolation step of the SEG-RR algorithm from (13) and let

$$\hat{z}_i^k = z_i^k - \gamma_2 F(z_i^k)$$

We start with the proof of the 1st point (inequality (5)).

Proof of Inequality (5). Using the update rule (14) of SEG-RR, we have that:

$$\begin{aligned} z_0^{k+1} \quad &= \quad z_n^k \\ &\overset{(14)}{=} \quad z_{n-1}^k - \gamma_1 F_{\pi_{n-1}^k}(\bar{z}_{n-1}^k) \\ &\overset{(14)}{=} \quad z_0^k - \gamma_1\sum_{i=0}^{n-1} F_{\pi_i^k}(\bar{z}_i^k) \\ &= \quad z_0^k - \gamma_1 nF(\hat{z}_0^k) \ - \gamma_1\sum_{i=0}^{n-1}\left(F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right) \end{aligned} \tag{43}$$

where we have expressed an epoch-level update by using (14) and in the last step we have added and subtracted the term $\gamma_1 nF(\hat{z}_0^k) = \gamma_1\sum\limits_{i=0}^{n-1} F_{\pi_i^k}(\hat{z}_0^k)$, utilizing the finite sum structure of the operator $F(z) = \frac{1}{n}\sum\limits_{i=0}^{n-1} F_{\pi_i^k}(z)$. Subtracting $z^\star$ from both sides of (43) and taking the norm, we get:

$$\|z_0^{k+1} - z_*\|^2 \quad = \quad \left\|z_0^k - z_* - \gamma_1 nF(\hat{z}_0^k) - \gamma_1\sum_{i=0}^{n-1}(F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}^k(\hat{z}_0^k))\right\|^2 \tag{44}$$

We, next, use Young's inequality (21) with $t = 1 - \frac{\gamma_1 n\mu}{2} \in [0,1]$ in order to expand the norm in the right-hand side (RHS) of (44) and then simplify the resulting terms. Specifically, we obtain:

$$\|z_0^{k+1} - z_*\|^2 \overset{(21)}{\leq} \frac{\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2}{1 - \frac{\gamma_1 n\mu}{2}} + \frac{2}{\gamma_1 n\mu}\left\|\gamma_1 \sum_{i=0}^{n-1}(F_i^k(\bar{z}_i^k) - F_i^k(\hat{z}_0^k))\right\|^2 \tag{45}$$

Taking expectation condition on the filtration $\mathcal{F}^k$ (history of z_0^k) and using Lemma A.5 to bound the second term in the right-hand side of (45) and get

$$\begin{aligned}\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \frac{1}{1 - \frac{\gamma_1 n\mu}{2}} \underbrace{\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2}_{T_1} + \frac{12\gamma_1 L_{max}^2}{\mu} \underbrace{\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|z_i^k - z_0^k\right\|^2 \middle| \mathcal{F}^k\right]}_{T_2} \\ &\quad + \frac{6\gamma_1\gamma_2^2 L_{max}^2}{\mu} \underbrace{\mathbb{E}\left[\sum_{i=0}^{n-1}\left\|F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]}_{T_3}\end{aligned} \tag{46}$$

We, next, use the upper bounds from Lemma B.1, A.4 and Proposition A.2 with $\gamma_2 = 2\gamma_1, \gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}$ in order to bound the terms T_1, T_2, T_3 as follows:

$$T_1 \quad \leq \quad \left(1 - \frac{1}{2}\gamma_1 n\mu\right)^2 \|z_0^k - z_*\|^2 + U\|z_0^k - z_*\|^2 \tag{47}$$

$$T_2 \quad \leq \quad \left[10n^2L^2 + A(25+n)\right] n\gamma_1^2 \left\|z_0^k - z_*\right\|^2 + 2n(n+25)\gamma_1^2\sigma_*^2 \tag{48}$$

$$T_3 \quad \leq \quad A\|z - z_*\|^2 + 2\sigma_*^2 \tag{49}$$

where $U = \left\{\gamma_1^2 n^2(2L^2 - \frac{\mu^2}{4}) + 2\gamma_1\gamma_2 nL^2\left[-1 + \gamma_2(\gamma_1 nL^2 + \gamma_2 L^2 + \mu)\right]\right\}$.

Substituting the upper bounds (47), (48), (49) into (46) and letting $C = 2\left[(25+n)A + 10n^2L^2\right]\gamma_1^2 + A\gamma_2^2$ for brevity, we get:

$$\begin{aligned}\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \left(1 - \frac{1}{2}\gamma_1 n\mu + \frac{U}{1 - \frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu}\right)\|z_0^k - z_*\|^2 \\ &\quad + \frac{24nL_{max}^2\gamma_1}{\mu}\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2\end{aligned} \tag{50}$$

Choosing the step size $\gamma_2 = 2\gamma_1, \gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$ appropriately and using Lemma (**??**), we can upper bound the term

$$\frac{U}{1 - \frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu} \quad \leq \quad \frac{\gamma_1 n\mu}{4} \tag{51}$$

Substituting (51) into (50), we obtain

$$\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad \left(1 - \frac{1}{4}\gamma_1 n\mu\right)\|z_0^k - z_*\|^2 + \frac{24nL_{max}^2\gamma_1\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{\mu}\sigma_*^2 \tag{52}$$

Taking expectation on both sides and using the tower property of expectations, we have that:

$$\begin{aligned}\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2\right] \quad &\leq \quad \left(1 - \frac{1}{4}\gamma_1 n\mu\right)\|z_0^k - z_*\|^2 + \frac{24nL_{max}^2\gamma_1\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{\mu}\sigma_*^2 \\ &\leq \quad \left(1 - \frac{1}{4}\gamma_1 n\mu\right)^{k+1}\|z_0^0 - z_*\|^2 + \frac{24nL_{max}^2\gamma_1\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{\mu}\sum_{i=1}^{k}(1 - \frac{1}{4}\gamma_1 n\mu)^i\sigma_*^2\end{aligned} \tag{53}$$

$$\begin{aligned}
&\leq \left(1-\frac{1}{4}\gamma_1 n\mu\right)^{k+1}\|z_0^0-z_*\|^2+\frac{24nL_{max}^2\gamma_1\left[(25+n)\gamma_1^2+\gamma_2^2\right]}{\mu}\sum_{i=1}^{\infty}(1-\frac{1}{4}\gamma_1 n\mu)^i\sigma_*^2\\
&= \left(1-\frac{\gamma_1 n\mu}{4}\right)^{k+1}\|z_0-z_*\|^2+\frac{96L_{max}^2}{\mu^2}\left[(25+n)\gamma_1^2+\gamma_2^2\right]\sigma_*^2 \qquad (54)
\end{aligned}$$

Proof of Equation (6). From (54) we have that

$$\begin{aligned}
\mathbb{E}\left[\|z_0^K-z_*\|^2\right] &\leq \left(1-\frac{\gamma_1 n\mu}{4}\right)^{K}\|z_0-z_*\|^2+\frac{96L_{max}^2}{\mu^2}\left[(25+n)\gamma_1^2+\gamma_2^2\right]\sigma_*^2\\
&\overset{\gamma_2=2\gamma_1}{\leq} \left(1-\frac{\gamma_1 n\mu}{4}\right)^{K}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^2}\gamma_1^2\sigma_*^2\\
&\overset{(19)}{\leq} e^{\frac{-\gamma_1 nK\mu}{4}}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^2}\gamma_1^2\sigma_*^2 \qquad (55)
\end{aligned}$$

We substitute $\gamma_1 = \min\left\{\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}, \frac{4\log(n^{1/2}K)}{\mu nK}\right\} \leq \frac{4\log(n^{1/2}K)}{\mu nK}$ and bound the second term in the right-hand side (RHS) of (55) as

$$\frac{96(29+n)L_{max}^2}{\mu^2}\gamma_1^2\sigma_*^2 \leq \frac{96(29+n)L_{max}^2}{\mu^4}\frac{\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (56)$$

Substituting (56) into (55), we obtain the following:

$$\mathbb{E}\left[\|z_0^{K+1}-z_*\|^2\right] \leq e^{-\frac{\gamma_1 nK\mu}{4}}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (57)$$

We now consider the following cases:

Case 1: $\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}} \leq \frac{\log(n^{1/2}K)}{\mu nK}$ In this case we have that $\gamma_1 = \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$, which implies that the RHS of (57) is bounded by

$$\begin{aligned}
& e^{-\frac{\gamma_1 nK\mu}{4}}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2\\
\leq\ & e^{-\frac{nK\mu^2}{40L_{max}^2\sqrt{10n^2+2n+54}}}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2\\
\leq\ & e^{-\frac{K\mu^2}{40\sqrt{12}L_{max}^2}}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 s \qquad (58)
\end{aligned}$$

Case 2: $\frac{4\log(n^{1/2}K)}{\mu nK} \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$ In this case we have that $\gamma_1 = \frac{4\log(n^{1/2}K)}{\mu nK}$, which implies that the RHS of (57) is bounded by

$$\begin{aligned}
& e^{-\frac{\gamma_1 nK\mu}{4}}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2\\
\leq\ & \frac{1}{nK^2}\|z_0-z_*\|^2+\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (59)
\end{aligned}$$

Taking the maximum of the right-hand side of (94) and (95) and using the inequality $\max\{a,b\} \leq a+b$, we obtain the desired result which holds for both cases:

$$\mathbb{E}\left[\|z_0^K-z_*\|^2\right] \leq e^{-\frac{K\mu^2}{40\sqrt{12}L_{max}^2}}\|z_0-z_*\|^2+\frac{1}{nK^2}\|z_0-z_*\|^2+2\frac{96(29+n)L_{max}^2}{\mu^4}\frac{16\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2$$

Suppressing constant and logarithmic terms, we get the final result

$$\mathbb{E}\left[\|z_0^{K+1}-z_*\|^2\right] = \tilde{\mathcal{O}}\left(e^{-\frac{K\mu^2}{L_{max}^2}}+\frac{1}{nK^2}\right)$$

□

B.2 Proofs for Affine Case

B.2.1 Lemma for Iterates in Affine Case

Lemma B.3. Suppose that $F_i, \forall i \in [n-1]$ are affine and Assumption 3 holds. If the step size of SEG-RR Algorithm satisfy $\gamma_2 = 4\gamma_1, \gamma_1 \in \left(0, \frac{1}{3\sqrt{2n(n-1)}L_{max}}\right]$ then the following holds

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I-\gamma_2 Q_i^k)b_i^k\right\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad 2nL_{max}D_2\gamma_1^2\left\|z_0^k - z_*\right\|^2 \\ &\quad +2L_{max}\left[\gamma_2^2 + 2n\gamma_1^2(n+25)\right]\sigma_*^2\end{aligned}$$

where $D_2 = \left[10n^2L^2 + (n+25)A\right]$.

Proof. We have that

$$\begin{aligned}&\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I-\gamma_2 Q_i^k)b_i^k\right\|^2 \\ = \quad &\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k) + \sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)z^* + (I-\gamma_2 Q_i^k)b_i^k\right\|^2 \\ = \quad &\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k) + (Qz^* + b) - \gamma_2\sum_{i=0}^{n-1} Q_i^k(Q_i^k z^* + b_i^k)\right\|^2 \\ \overset{(7)}{=} \quad &\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k) + F(z^*) - \gamma_2\sum_{i=0}^{n-1} Q_i^k F_i^k(z^*)\right\|^2\end{aligned}$$

Using the fact that $F(z_*) = 0$, we continue our derivation as follows

$$\begin{aligned}&\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I-\gamma_2 Q_i^k)b_i^k\right\|^2 \\ = \quad &\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k) - \gamma_2\sum_{i=0}^{n-1} Q_i^k F_i^k(z^*)\right\|^2 \\ \overset{(16)}{\leq} \quad &2\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k)\right\|^2 + 2\left\|\gamma_2\sum_{i=0}^{n-1} Q_i^k F_i^k(z^*)\right\|^2 \\ \overset{(16)}{\leq} \quad &2n\sum_{i=0}^{n-1}\left\|Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k)\right\|^2 + 2\gamma_2^2\sum_{i=0}^{n-1}\left\|Q_i^k F_i^k(z^*)\right\|^2 \\ \leq \quad &2n\left(L_{max} - \gamma_2\lambda_{min}(Q_i^k)\right)\sum_{i=0}^{n-1}\left\|(z_i^k - z_0^k)\right\|^2 + 2L_{max}\gamma_2^2\sum_{i=0}^{n-1}\left\|F_i^k(z^*)\right\|^2 \\ = \quad &2n\left(L_{max} - \gamma_2\lambda_{min}(Q_i^k)\right)\sum_{i=0}^{n-1}\left\|(z_i^k - z_0^k)\right\|^2 + 2L_{max}\gamma_2^2\sigma_*^2 \qquad (60)\end{aligned}$$

where $\sigma_*^2 = \frac{1}{n}\sum_{i=1}^n \left\|F_i^k(z^*)\right\|^2$ and $\lambda_{min}(Q_i^k) = \min_{i\in[n]}\min_{\lambda}\lambda(Q_i)$ is the minimum eigenvalue of all $Q_i^k, i \in [n]$. Taking expectation on both sides condition on the filtration $\mathcal{F}^k$, we get

$$\mathbb{E}\left[\left\|\sum_{i=0}^{n-1} Q_i^k(I-\gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I-\gamma_2 Q_i^k)b_i^k\right\|^2 \middle| \mathcal{F}^k\right]$$

$$\leq \quad 2n\left(L_{max} - \gamma_2 \lambda_{min}(Q_i^k)\right) \mathbb{E}\left[\sum_{i=0}^{n-1} \left\|(z_i^k - z_0^k)\right\|^2 \Big| \mathcal{F}^k\right] + 2L_{max}\gamma_2^2\sigma_*^2$$

Using lemma A.4, we continue our derivation as follows:

$$\begin{aligned}
& \mathbb{E}\left[\left\|\sum_{i=0}^{n-1} Q_i^k(I - \gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I - \gamma_2 Q_i^k)b_i^k\right\|^2 \Big| \mathcal{F}^k\right] \\
\overset{\text{Lemma } A.4}{\leq} \quad & 2n\left(L_{max} - \gamma_2\lambda_{min}(Q_i^k)\right)\left\{10n^2\gamma_1^2\left\|F(z_0^k)\right\|^2 + A\gamma_1^2(25+n)\left\|z_0^k - z_*\right\|^2\right\} \\
& + 4n\left(L_{max} - \gamma_2\lambda_{min}(Q_i^k)\right)(n+25)\gamma_1^2\sigma_*^2 + 2L_{max}\gamma_2^2\sigma_*^2 \\
\leq \quad & 2nL_{max}\left\{10n^2\gamma_1^2\left\|F(z_0^k)\right\|^2 + A\gamma_1^2(25+n)\left\|z_0^k - z_*\right\|^2\right\} \\
& + 2L_{max}\left[\gamma_2^2 + 2n(n+25)\gamma_1^2\right]\sigma_*^2
\end{aligned}$$

Lastly, applying the Lipschitz property of F (Assumption 3), we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=0}^{n-1} Q_i^k(I - \gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I - \gamma_2 Q_i^k)b_i^k\right\|^2 \Big| \mathcal{F}^k\right] \overset{(3)}{\leq} \quad & 2nL_{max}\gamma_1^2\left[10n^2L^2 + (n+25)A\right]\left\|z_0^k - z_*\right\|^2 \\
& + 2L_{max}\left[\gamma_2^2 + 2n\gamma_1^2(n+25)\right]\sigma_*^2
\end{aligned}$$

□

Lemma B.4. If $\gamma_2 = 4\gamma_1$, $\gamma_1 \in \left(0, \frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}}\right]$, then the following hold:

$$1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max}) + \frac{2L_{max}\gamma_1^3}{\lambda^+_{min}(Q)}\left[10n^2L^2 + (n+25)A\right] \quad \leq \quad 1 - \frac{\gamma_1 n \lambda^+_{min}(Q)}{2}$$

Proof. Selecting the step size $\gamma_1, \gamma_2 = 4\gamma_1$ such that

$$\begin{aligned}
& 1 - \gamma_1 n(\lambda^+_{min}(Q) - 4\gamma_1 L^2_{max}) + \frac{2L_{max}\gamma_1^3}{\lambda^+_{min}(Q)}\left[10n^2L^2 + (n+25)A\right] \leq 1 - \frac{\gamma_1 n\lambda^+_{min}(Q)}{2} \\
\iff \quad & \lambda^+_{min}(Q) - 4\gamma_1 L^2_{max} - \frac{2L_{max}\gamma_1^2}{\lambda^+_{min}(Q)n}\left[10n^2L^2 + (n+25)A\right] \geq \frac{\lambda^+_{min}(Q)}{2} \\
\iff \quad & \frac{\lambda^+_{min}(Q)}{2} - 4L^2_{max}\gamma_1 - \frac{2L_{max}\gamma_1^2}{\lambda^+_{min}(Q)n}\left[10n^2L^2 + (n+25)A\right] \geq 0
\end{aligned} \tag{61}$$

Using the fact that $A = \frac{2}{n}\sum\limits_{i=0}^{n-1} L_i^2 \leq 2L^2_{max}$, it suffice to select the stepsize γ_1 such that

$$\frac{\lambda^+_{min}(Q)}{2} - 4L^2_{max}\gamma_1 - \frac{2L^3_{max}\gamma_1^2}{\lambda^+_{min}(Q)n}(5n^2 + 2n + 50) \geq 0 \tag{62}$$

In order to get simple expressions for the step size instead of solving the quadratic inequality (62) we select γ_1 such that

$$\begin{aligned}
& \frac{\lambda^+_{min}(Q)}{4} - 4L^2_{max}\gamma_1 \geq 0 \quad \text{and} \quad \frac{\lambda^+_{min}(Q)}{4} - \frac{2L^3_{max}\gamma_1^2}{\lambda^+_{min}(Q)n}(5n^2 + 2n + 50) \geq 0 \\
\iff \quad & \gamma_1 \leq \frac{\lambda^+_{min}(Q)}{16L^2_{max}} \quad \text{and} \quad \gamma_1 \leq \frac{\lambda^+_{min}(Q)}{2L_{max}\sqrt{2L_{max}(5n + 2 + \frac{50}{n})}}
\end{aligned}$$

Thus, the above two constraints are satisfied for stepsize

$$\gamma_1 \leq \frac{\lambda^+_{min}(Q)}{2L^2_{max}\sqrt{2(5n+52)}}$$

Combining, lastly, the requirement that $\gamma_1 \leq \frac{1}{3L_{max}\sqrt{2n(n-1)}}$, we have that

$$\begin{aligned}
\gamma_1 &= \min\left\{\frac{\lambda^+_{min}(Q)}{2L^2_{max}\sqrt{2(5n+52)}}, \frac{1}{3L_{max}\sqrt{2n(n-1)}}\right\} \\
&\leq \min\left\{\frac{\lambda^+_{min}(Q)}{2L^2_{max}\sqrt{120n}}, \frac{1}{3\sqrt{2}nL_{max}}\right\} \\
&\leq \frac{\lambda^+_{min}(Q)}{2\sqrt{120n}L^2_{max}}
\end{aligned}$$

Thus, in order for the inequality in the statement of the Lemma to hold, it suffices to select the step size $\gamma_2 = 4\gamma_1, \gamma_1 \in (0, \gamma_{1,max}]$, where $\gamma_{1,max} = \frac{\lambda^+_{min}(Q)}{2\sqrt{120n}L^2_{max}}$. □

B.2.2 Proof of Theorem 2.2

Proof. Let the step size satisfy $\gamma_2 = \alpha\gamma_1, \alpha > 0$. We note that, due to the closed form expression (7) of the operator F, we have that the following hold

$$F^k_i(\bar{z}^k_i) = Q^k_i(I - \alpha\gamma_1 Q^k_i)z^k_i + (I - \alpha\gamma_1 Q^k_i)b^k_i \tag{63}$$
$$F(\hat{z}) = Q(I - \alpha\gamma_1 Q)z + (I - \alpha\gamma_1 Q)b \tag{64}$$
$$F(z_*) = 0 \iff Qz_* = -b \tag{65}$$

Proof of Inequality (8). We have that:

$$\begin{aligned}
z^{k+1}_0 &:= z^k_n \\
&\overset{(13)}{=} z^k_{n-1} - \gamma_1 F_{\pi^k_{n-1}}(\bar{z}^k_{n-1}) \\
&= z^k_0 - \gamma_1 \sum_{i=0}^{n-1} F_{\pi^k_i}(\bar{z}^k_i)
\end{aligned} \tag{66}$$

Subtracting $z^\star$ from both sides of (66) and taking the norm, we get:

$$\begin{aligned}
\|z^{k+1}_0 - z_*\|^2 &= \left\| z^k_0 - z_* - \gamma_1 \sum_{i=0}^{n-1} F_{\pi^k_i}(\bar{z}^k_i) \right\|^2 \\
&\overset{(63)}{=} \left\| z^k_0 - z_* - \gamma_1 \sum_{i=0}^{n-1} \left[Q^k_i(I - \gamma_2 Q^k_i)z^k_i + (I - \gamma_2 Q^k_i)b^k_i\right] \right\|^2 \\
&= \Bigg\| \left[I - \gamma_1 \sum_{i=0}^{n-1} Q^k_i(I - \gamma_2 Q^k_i)\right](z^k_0 - z_*) \\
&\quad -\gamma_1 \sum_{i=0}^{n-1} \left[Q^k_i(I - \gamma_2 Q^k_i)(z^k_i - z^k_0 + z^*) + (I - \gamma_2 Q^k_i)b^k_i\right] \Bigg\|^2
\end{aligned}$$

where in the last step we have added and subtracted the term $\gamma_1 \sum_{i=0}^{n-1} Q^k_i(I - \gamma_2 Q^k_i)(z^k_0 - z_*)$.

Using Young's inequality (22) with $t = 1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max}) \in (0,1)$, we obtain:

$$\begin{aligned}
\|z_0^{k+1} - z_*\|^2 &\overset{(22)}{\leq} \frac{\left\|I - \gamma_1 \sum_{i=0}^{n-1} Q_i^k (I - \gamma_2 Q_i^k)\right\|^2 \|z_0^k - z_*\|^2}{[1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})]} \\
&\quad + \frac{\left\|\gamma_1 \sum_{i=0}^{n-1} \left[Q_i^k (I - \gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I - \gamma_2 Q_i^k) b_i^k\right]\right\|^2}{\gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})} \\
&\leq [1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})]\|z_0^k - z_*\|^2 \\
&\quad + \frac{\gamma_1 \left\|\sum_{i=0}^{n-1} Q_i^k (I - \gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I - \gamma_2 Q_i^k) b_i^k\right\|^2}{n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})}
\end{aligned}$$

Taking expectation condition on the filtration $\mathcal{F}^k$, we have that

$$\begin{aligned}
\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] &\leq [1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})]\|z_0^k - z_*\|^2 \\
&\quad + \frac{\gamma_1 \mathbb{E}\left[\left\|\sum_{i=0}^{n-1} Q_i^k (I - \gamma_2 Q_i^k)(z_i^k - z_0^k + z^*) + (I - \gamma_2 Q_i^k) b_i^k\right\|^2 \middle| \mathcal{F}^k\right]}{n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})}
\end{aligned} \tag{67}$$

We, next, use Lemma B.3 to bound the second norm in the RHS of (67). Thus, letting $\gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}$ and using Lemma B.3 into (67) we get

$$\begin{aligned}
\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] &\overset{\text{Lemma } B.3}{\leq} \left[1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max}) + \frac{2L_{max}(10n^2L^2 + (n+25)A)\gamma_1^3}{(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})}\right]\|z_0^k - z_*\|^2 \\
&\quad + \frac{2L_{max}\gamma_1}{n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})}\left[\gamma_2^2 + 2n\gamma_1^2(n+25)\right]\sigma_*^2
\end{aligned} \tag{68}$$

Selecting the step size $\gamma_2 = 4\gamma_1$, $\gamma_1 \leq \frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}}$ and using Lemma B.4 we have that:

$$[1 - \gamma_1 n(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})] + \frac{2L_{max}\gamma_1^3}{(\lambda^+_{min}(Q) - \gamma_2 L^2_{max})}\left[10n^2L^2 + (n+25)A\right] \leq 1 - \frac{\gamma_1 n \lambda^+_{min}(Q)}{2}$$

Thus, for the selected step size inequality (68) gives:

$$\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \leq \left(1 - \frac{1}{2}\gamma_1 n \lambda^+_{min}(Q)\right)\|z_0^k - z_*\|^2 + \frac{2L_{max}\gamma_1}{n\lambda^+_{min}(Q)}\left[\gamma_2^2 + 4n\gamma_1^2(n+25)\right]\sigma_*^2 \tag{69}$$

Taking expectation on both sides and using the tower property of expectation we have that

$$\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2\right] \leq \left(1 - \frac{1}{2}\gamma_1 n \lambda^+_{min}(Q)\right)\|z_0^k - z_*\|^2 + \frac{2L_{max}\gamma_1}{n\lambda^+_{min}(Q)}\left[\gamma_2^2 + 4n\gamma_1^2(n+25)\right]\sigma_*^2 \tag{70}$$

Using the concavity of the min operator and the definition of $\text{dist}(z, \mathcal{Z}_*) = \min_{z_* \in \mathcal{Z}_*} \|z - z_*\|^2$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\text{dist}(z_0^{k+1}, \mathcal{Z}_*)\right] &\leq \min_{z_* \in \mathcal{Z}_*} \mathbb{E}\left[\|z_0^{k+1} - z_*\|^2\right] \\
&\leq \left(1 - \frac{1}{2}\gamma_1 n \lambda^+_{min}(Q)\right)\text{dist}(z_0^k, \mathcal{Z}_*) + \frac{2L_{max}\gamma_1}{n\lambda^+_{min}(Q)}\left[\gamma_2^2 + 4n\gamma_1^2(n+25)\right]\sigma_*^2
\end{aligned} \tag{71}$$

where σ_*^2 here denotes $\sigma_*^2 = \min_{z_* \in \mathcal{Z}_*} \frac{1}{n}\sum_{i=0}^{n-1}\|F_i(z_*)\|^2$.
Unrolling the recursion, we conclude that

$$\mathbb{E}\left[\text{dist}(z_0^{k+1}, \mathcal{Z}_*)\right] \leq \left(1 - \frac{1}{2}\gamma_1 n \lambda^+_{min}(Q)\right)^{k+1}\text{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}}{\lambda^2_{min}(Q)n^2}\left[\gamma_2^2 + 4n(n+25)\gamma_1^2\right]\sigma_*^2 \tag{72}$$

Proof of Equation (9). From inequality (72), the following holds

$$\begin{aligned}
\mathbb{E}\left[\operatorname{dist}(z_0^K, \mathcal{Z}_*)\right] \quad &\leq \quad \left(1 - \frac{1}{2}\gamma_1 n \lambda^+_{min}(Q)\right)^K \operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}}{\lambda^2_{min}(Q)n^2}\left[\gamma_2^2 + 4n(n+25)\gamma_1^2\right]\sigma_*^2 \\
&\overset{\gamma_2 = 4\gamma_1}{\leq} \quad \left(1 - \frac{1}{2}\gamma_1 n \lambda^+_{min}(Q)\right)^K \operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}}{\lambda^2_{min}(Q)n^2}\left(4n^2 + 25n + 16\right)\gamma_1^2\sigma_*^2 \\
&\overset{(19)}{\leq} \quad e^{-\frac{\gamma_1 n K \lambda^+_{min}(Q)}{2}}\operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^2_{min}(Q)n^2}\gamma_1^2\sigma_*^2 \qquad (73)
\end{aligned}$$

We substitute $\gamma_1 = \min\left\{\frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}}, \frac{2\log(n^{1/2}K)}{\lambda^+_{min}(Q)nK}\right\} \leq \frac{2\log(n^{1/2}K)}{\lambda^+_{min}(Q)nK}$ and bound the second term in the right-hand side (RHS) of (73) as

$$\frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^2_{min}(Q)n^2}\gamma_1^2\sigma_*^2 \leq \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (74)$$

Substituting (74) into (73), we obtain the following:

$$\mathbb{E}\left[\operatorname{dist}(z_0^K, \mathcal{Z}_*)\right] \leq e^{-\frac{\gamma_1 n K \lambda^+_{min}(Q)}{2}}\operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (75)$$

We now consider the following cases:

Case 1: $\frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}} \leq \frac{2\log(n^{1/2}K)}{\lambda^+_{min}(Q)nK}$ In this case we have that $\gamma_1 = \frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}}$, which implies that the RHS of (75) is bounded by

$$\begin{aligned}
&e^{-\frac{\gamma_1 n K \lambda^+_{min}(Q)}{2}}\operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \\
\leq \quad &e^{-\frac{K\lambda^2_{min}(Q)}{4\sqrt{120}L^2_{max}}}\|z_0 - z_*\|^2 + \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (76)
\end{aligned}$$

Case 2: $\frac{2\log(n^{1/2}K)}{\lambda^+_{min}(Q)nK} \leq \frac{\lambda^+_{min}(Q)}{2\sqrt{120}nL^2_{max}}$ In this case we have that $\gamma_1 = \frac{2\log(n^{1/2}K)}{\lambda^+_{min}(Q)nK}$, which implies that the RHS of (75) is bounded by

$$\begin{aligned}
&e^{-\frac{\gamma_1 n K \lambda^+_{min}(Q)}{2}}\operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \\
\leq \quad &\frac{1}{nK^2}\|z_0 - z_*\|^2 + \frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2 \qquad (77)
\end{aligned}$$

Taking the maximum of the right-hand side of (76) and (77) and using the inequality $\max\{a, b\} \leq a + b$, we obtain the desired result which holds for both cases:

$$\mathbb{E}\left[\operatorname{dist}(z_0^K, \mathcal{Z}_*)\right] \leq e^{-\frac{K\lambda^2_{min}(Q)}{4\sqrt{120}L^2_{max}}}\operatorname{dist}(z_0, \mathcal{Z}_*) + \frac{1}{nK^2}\|z_0 - z_*\|^2 + 2\frac{4L_{max}\left(4n^2 + 25n + 16\right)}{\lambda^+_{min}(Q)^4 n^2}\frac{4\log^2(n^{1/2}K)}{n^2K^2}\sigma_*^2$$

Suppressing constant and logarithmic terms, we get the final result

$$\mathbb{E}\left[\operatorname{dist}(z_0^K, \mathcal{Z}_*)\right] \quad = \quad \tilde{\mathcal{O}}\left(e^{\frac{K\lambda^2_{min}(Q)}{4\sqrt{120}L^2_{max}}} + \frac{1}{nK^2}\right)$$

□

B.3 Proofs for Monotone Case

B.3.1 Lemma for Iterates in Monotone Case

We start with a lemma bounding the iterates in the monotone case when the full-batch operator F is used.

Lemma B.5. Suppose that the operator F is monotone and each F_i, $\forall i \in [n]$ is L_i-Lipschitz. If SEG-RR is run with extrapolation stepsize $\gamma_2 \leq \frac{1}{L}$, then the iterates of SEG-RR satisfy

$$\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 \quad \leq \quad \left(1 + 4\gamma_1^2 n^2 L^2\right)\|z_0^k - z_*\|^2 - \gamma_1\gamma_2 n(1-\gamma_2^2 L^2)\|F(z_0^k)\|^2$$

Proof.

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &= \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle z_0^k - z_*, F(\hat{z}_0^k)\rangle + \gamma_1^2 n^2\|F(\hat{z}_0^k)\|^2 \\
&= \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle + \gamma_1^2 n^2\|F(\hat{z}_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(13)}{=} \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad +\gamma_1^2 n^2\|F(\hat{z}_0^k) - F(z_0^k) + F(z_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(16)}{=} \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad +2\gamma_1^2 n^2\|F(\hat{z}_0^k) - F(z_0^k)\|^2 + 2\gamma_1^2 n^2\|F(z_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle
\end{aligned}$$

Using the Lipschitz property of the operator F, we get that

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &\overset{(3)}{\leq} \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad +2\gamma_1^2 n^2 L^2\|\hat{z}_0^k - z_0^k\|^2 + 2\gamma_1^2 n^2\|F(z_0^k)\|^2 \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(15)}{=} \|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad +2\gamma_1^2 n^2(1+\gamma_2^2 L^2)\|F(z_0^k)\|^2 - 2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle \\
&\overset{(3)}{\leq} [1 + 2\gamma_1^2 n^2 L^2(1+\gamma_2^2 L^2)]\|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad -2\gamma_1\gamma_2 n\langle F(z_0^k), F(\hat{z}_0^k)\rangle
\end{aligned}$$

We continue with the use of inequality (18):

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &\overset{(18)}{\leq} [1 + 2\gamma_1^2 n^2 L^2(1+\gamma_2^2 L^2)]\|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad -\gamma_1\gamma_2 n\left\|F(z_0^k)\right\|^2 - \gamma_1\gamma_2 n\left\|F(\hat{z}_0^k)\right\|^2 + \gamma_1\gamma_2 n\left\|F(\hat{z}_0^k) - F(z_0^k)\right\|^2 \\
&\overset{(3)}{\leq} [1 + 2\gamma_1^2 n^2 L^2(1+\gamma_2^2 L^2)]\|z_0^k - z_*\|^2 - 2\gamma_1 n\langle \hat{z}_0^k - z_*, F(\hat{z}_0^k)\rangle \\
&\quad -\gamma_1\gamma_2 n(1-\gamma_2^2 L^2)\|F(z_0^k)\|^2
\end{aligned}$$

Using, as a last step, the fact that the operator F is monotone and $\gamma_2 \leq \frac{1}{L}$, we get

$$\begin{aligned}
\left\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 &\leq \left[1 + 2\gamma_1^2 n^2 L^2(1+\gamma_2^2 L^2)\right]\|z_0^k - z_*\|^2 - \gamma_1\gamma_2 n(1-\gamma_2^2 L^2)\|F(z_0^k)\|^2 \\
&\overset{\gamma_2 \leq \frac{1}{L}}{\leq} \left(1 + 4\gamma_1^2 n^2 L^2\right)\|z_0^k - z_*\|^2 - \gamma_1\gamma_2 n(1-\gamma_2^2 L^2)\|F(z_0^k)\|^2
\end{aligned}$$

□

B.3.2 Proof of Theorem 2.3

Proof. We start with the proof of the first point (inequality (10)) in the statement of the Theorem 2.3.

Proof of Inequality (10). Using the update rule in (13), we have that:

$$\begin{aligned} z_0^{k+1} &= z_n^k \\ &\overset{(13)}{=} z_{n-1}^k - \gamma_1 F_{n-1}^k(\bar{z}_{n-1}^k) \\ &= z_0^k - \gamma_1 \sum_{i=0}^{n-1} F_{\pi_i^k}(\bar{z}_i^k) \\ &= z_0^k - \gamma_1 n F(\hat{z}_0^k) \; - \gamma_1 \sum_{i=0}^{n-1} \left(F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right) \end{aligned} \tag{78}$$

where in the last step we add and subtract the term $\gamma_1 n F(\hat{z}_0^k) = \gamma_1 \sum_{i=0}^{n-1} F_{\pi_i^k}(\hat{z}_0^k)$.

Subtracting $z^\star$ from both sides of (78) and taking the norm, we get

$$\begin{aligned} \|z_0^{k+1} - z_*\|^2 &= \left\| z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k) - \gamma_1 \sum_{i=0}^{n-1} \left(F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right) \right\|^2 \\ &\overset{(16)}{\leq} 2\|z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\|^2 + 2\gamma_1^2 \left\| \sum_{i=0}^{n-1} F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k) \right\|^2 \\ &\overset{(16)}{\leq} 2\left\| z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k) \right\|^2 + 2\gamma_1^2 n \sum_{i=0}^{n-1} \left\| F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k) \right\|^2 \end{aligned} \tag{79}$$

Taking expectation on both sides condition on the filtration $\mathcal{F}_k$ and applying Lemma A.5 results to

$$\begin{aligned} \mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] &\leq 2\mathbb{E}\left[\left\| z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 \middle| \mathcal{F}^k\right] + 2\gamma_1^2 n \mathbb{E}\left[\sum_{i=0}^{n-1} \left\| F_{\pi_i^k}(\bar{z}_i^k) - F_{\pi_i^k}(\hat{z}_0^k)\right\|^2 \middle| \mathcal{F}^k\right] \\ &\overset{\text{Lemma } A.5}{\leq} 2\underbrace{\mathbb{E}\left[\left\| z_0^k - z_* - \gamma_1 n F(\hat{z}_0^k)\right\|^2 \middle| \mathcal{F}^k\right]}_{T_1} + 12 n L_{max}^2 \gamma_1^2 \underbrace{\mathbb{E}\left[\sum_{i=0}^{n-1} \left\| z_i^k - z_0^k \right\|^2 \middle| \mathcal{F}^k\right]}_{T_2} \\ &\quad + 6 n \gamma_1^2 \gamma_2^2 \underbrace{\mathbb{E}\left[\sum_{i=0}^{n-1} \left\| F_{\pi_i^k}(z_0^k) - F(z_0^k)\right\|^2 \middle| \mathcal{F}^k\right]}_{T_3} \end{aligned} \tag{80}$$

Next, we bound the terms T_1, T_2, T_3 in (80) using Lemma B.5, A.4 and Proposition A.5 for $\gamma_2 \leq \frac{1}{L}$, as follows:

$$\begin{aligned} T_1 &\leq 2(1 + 4\gamma_1^2 n^2 L^2)\|z_0^k - z_*\|^2 - 2\gamma_1\gamma_2 n(1 - \gamma_2^2 L^2)\|F(z_0^k)\|^2 \\ T_2 &\leq \left[10 n^2 L^2 + A(25+n)\right] n\gamma_1^2 \left\| z_0^k - z_* \right\|^2 + 2n(n+25)\gamma_1^2 \sigma_*^2 \end{aligned} \tag{81}$$

$$T_3 \leq A\|z - z_*\|^2 + 2\sigma_*^2 \tag{82}$$

Substituting (81), (81), (82) into (80), we get

$$\begin{aligned} \mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] &\leq 2(1 + 4\gamma_1^2 n^2 L^2)\|z_0^k - z_*\|^2 - 2\gamma_1\gamma_2 n(1 - \gamma_2^2 L^2)\|F(z_0^k)\|^2 \\ &\quad + 6\gamma_1^2 n^2 C L_{max}^2 \|z_0^k - z_*\|^2 + 24\gamma_1^2 n^2 L_{max}^2 \left[(25+n)\gamma_1^2 + \gamma_2^2\right] \sigma_*^2 \\ &= 2(1 + 4\gamma_1^2 n^2 L^2 + 3C L_{max}^2)\|z_0^k - z_*\|^2 - 2\gamma_1\gamma_2 n(1 - \gamma_2^2 L^2)\|F(z_0^k)\|^2 \\ &\quad + 24\gamma_1^2 n^2 L_{max}^2 \left[(25+n)\gamma_1^2 + \gamma_2^2\right] \sigma_*^2 \end{aligned} \tag{83}$$

where $C = 2\left[(25+n)A + 10n^2L^2\right]\gamma_1^2 + A\gamma_2^2$.
For $\gamma_2 = 2\gamma_1$ and $\gamma_1 \leq \frac{1}{3\sqrt{2}nL_{max}}$, we have that $C \leq \frac{2A+10L^2}{9L_{max}^2}$ and thus

$$(1 + 4\gamma_1^2 n^2 L^2 + 3CL_{max}^2) \quad \leq \quad 3(1 + CL_{max}^2) \overset{C \leq \frac{2A+10L^2}{9L_{max}^2}}{\leq} 3\left(A + 4L^2 + 1\right)$$

Let $G = 6\left(A + 4L^2 + 1\right)$. Rearranging the terms in (83) and taking expectation condition on $\mathcal{F}^k$, we have that

$$\begin{aligned} 2\gamma_1\gamma_2 n(1-\gamma_2^2L^2)\mathbb{E}\left[\|F(z_0^k)\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \mathbb{E}\left[G\|z_0^k - z_*\|^2 - \|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \\ &\quad + 24\gamma_1^2 n^2 L_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2 \end{aligned}$$

For $\gamma_2 \leq \frac{1}{L}$, we have that $(1-\gamma_2^2L^2) \geq 0$ and thus obtain

$$\begin{aligned} \mathbb{E}\left[\|F(z_0^k)\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \frac{\mathbb{E}\left[G\|z_0^k - z_*\|^2 - \|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right]}{2\gamma_1\gamma_2 n(1-\gamma_2^2L^2)} + \frac{12\gamma_1 n L_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{\gamma_2(1-\gamma_2^2L^2)}\sigma_*^2 \\ \overset{\gamma_2=2\gamma_1}{\Longleftrightarrow} \mathbb{E}\left[\|F(z_0^k)\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \frac{\mathbb{E}\left[G\|z_0^k - z_*\|^2 - \|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right]}{4\gamma_1^2 n(1-\gamma_2^2L^2)} + \frac{6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{(1-\gamma_2^2L^2)}\sigma_*^2 \qquad (84) \end{aligned}$$

Taking expectation on both sides and using the tower law of expectation, we get that:

$$\begin{aligned} \mathbb{E}\left[\|F(z_0^k)\|^2\right] \quad &\leq \quad \frac{1}{4\gamma_1^2 n(1-\gamma_2^2L^2)}\mathbb{E}\left[G\|z_0^k - z_*\|^2 - \|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \\ &\quad + \frac{6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{(1-\gamma_2^2L^2)}\sigma_*^2 \qquad (85) \end{aligned}$$

Let $G_k = \left(\frac{1}{G}\right)^k$ be a sequence of k. Multiplying both sides of the above inequality by G_k results to

$$\begin{aligned} G_k\mathbb{E}\left[\|F(z_0^k)\|^2\right] \quad &\leq \quad \frac{1}{4\gamma_1^2 n(1-\gamma_2^2L^2)}\mathbb{E}\left[G_{k-1}\|z_0^k - z_*\|^2 - G_k\|z_0^{k+1} - z_*\|^2\right] \\ &\quad + \frac{6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{(1-\gamma_2^2L^2)}G_k\sigma_*^2 \end{aligned}$$

Using the fact that $G_k \leq 1, \forall k \geq 0$, we have that

$$\begin{aligned} G_k\mathbb{E}\left[\|F(z_0^k)\|^2\right] \quad &\leq \quad \frac{1}{4\gamma_1^2 n(1-\gamma_2^2L^2)}\mathbb{E}\left[G_{k-1}\|z_0^k - z_*\|^2 - G_k\|z_0^{k+1} - z_*\|^2\right] \\ &\quad + \frac{6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{(1-\gamma_2^2L^2)}\sigma_*^2 \end{aligned}$$

Summing for $k = 0, ..., K$ and dividing by K, we obtain

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K} G_k\mathbb{E}\left[\|F(z_0^k)\|^2\right] \quad &\leq \quad \frac{1}{4\gamma_1^2 n(1-\gamma_2^2L^2)K}\sum_{k=0}^{K}\mathbb{E}\left[G_{k-1}\|z_0^k - z_*\|^2 - G_k\|z_0^{k+1} - z_*\|^2\right] \\ &\quad + \frac{6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]}{(1-\gamma_2^2L^2)}\sigma_*^2 \end{aligned}$$

Using Jensen inequality and letting $\tilde{z}_0^k = \frac{1}{K}\sum\limits_{k=0}^{K} G_k z_0^k$, we get the final result

$$\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right] \quad \leq \quad \frac{\|z_0 - z_*\|^2}{4nG\gamma_1^2(1-\gamma_2^2L^2)K} + \frac{6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2}{(1-\gamma_2^2L^2)} \qquad (86)$$

Using the fact that $\gamma_1 \leq \frac{1}{3nL_{max}}$ and $\frac{1}{(1-\gamma_2^2L^2)} = \frac{1}{(1-4\gamma_1^2L^2)} \leq \frac{9n-4}{9n} \leq 1$, we can simplify the above expression into the following

$$\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right] \quad \leq \quad \frac{\|z_0 - z_*\|^2}{4nG\gamma_1^2 K} + 6nL_{max}^2\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2$$

Proof of Equation (11)**.** From (72) we have that

$$\begin{aligned}\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right] &\leq \frac{\|z_0-z_*\|^2}{4nG\gamma_1^2K}+6nL_{max}^2\left[(25+n)\gamma_1^2+\gamma_2^2\right]\sigma_*^2\\ \overset{\gamma_2=2\gamma_1}{\iff} \mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right] &\leq \frac{\|z_0-z_*\|^2}{4nG\gamma_1^2K}+6nL_{max}^2\left(29+n\right)\gamma_1^2\sigma_*^2 \qquad (87)\end{aligned}$$

We substitute $\gamma_1=\min\left\{\frac{1}{3\sqrt{2}nL_{max}},\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right\}\leq\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}$ and bound the second term in the right-hand side (RHS) of (87) as

$$6nL_{max}^2\left(29+n\right)\gamma_1^2\sigma_*^2\leq\frac{6n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}} \qquad (88)$$

Substituting (88) into (87), we obtain the following:

$$\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right]\leq\frac{\|z_0-z_*\|^2}{4nG\gamma_1^2K}+\frac{6n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}} \qquad (89)$$

We now consider the following cases:

Case 1: $\frac{1}{3\sqrt{2}nL_{max}}\leq\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}$ In this case, we have that $\gamma_1=\frac{1}{3\sqrt{2}nL_{max}}$, which implies that the RHS of (89) is bounded by

$$\frac{\|z_0-z_*\|^2}{4nG\gamma_1^2K}+\frac{6n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}}\quad\leq\quad\frac{9nL_{max}^2\|z_0-z_*\|^2}{2GK}+\frac{6n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}} \qquad (90)$$

Case 2: $\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\leq\frac{1}{3\sqrt{2}nL_{max}}$ In this case we have that $\gamma_1=\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}$, which implies that the RHS of (89) is bounded by

$$\frac{\|z_0-z_*\|^2}{4nG\gamma_1^2K}+\frac{6n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}}\quad\leq\quad\frac{\|z_0-z_*\|^2}{4Gn^{\frac{1}{3}}K^{\frac{1}{3}}}+\frac{6n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}} \qquad (91)$$

Taking the maximum of the right-hand side of (90) and (91) and using the inequality $\max\{a,b\}\leq a+b$, we obtain the desired result which holds for both cases:

$$\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right]\leq\frac{9nL_{max}^2\|z_0-z_*\|^2}{2GK}+\frac{\|z_0-z_*\|^2}{4Gn^{\frac{1}{3}}K^{\frac{1}{3}}}+\frac{12n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}}$$

Lastly, we note that after a number of epochs the second term will dominate in the above inequality and the rate of convergence will be $\mathcal{O}\left(\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right)$. To show the aforementioned convergence rate, one can find a large enough constant (i.e. $C=60L_{max}^2\max\{\|z_0-z_*\|^2,\sigma_*^2\}$) such that it holds

$$\begin{aligned}\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right] &\leq \frac{9nL_{max}^2\|z_0-z_*\|^2}{2GK}+\frac{\|z_0-z_*\|^2}{4Gn^{\frac{1}{3}}K^{\frac{1}{3}}}+\frac{12n^{\frac{1}{3}}L_{max}^2\left(29+n\right)\sigma_*^2}{K^{\frac{2}{3}}}\leq C\left(\frac{n}{K}+\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}+\frac{n^{\frac{4}{3}}}{K^{\frac{2}{3}}}\right)\\ &\leq C\left(\frac{2n^{\frac{4}{3}}}{K}+\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right)\end{aligned}$$

and find the number of epochs K such that the following holds

$$\begin{aligned}C\left(\frac{2n^{\frac{4}{3}}}{K}+\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right) &\leq \frac{2C}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\\ \iff C\frac{2n^{\frac{4}{3}}}{K} &\leq \frac{C}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\\ \iff K &\geq 2^{\frac{3}{2}}n^{\frac{5}{2}} \qquad (92)\end{aligned}$$

Hence, after $K\geq\mathcal{O}\left(n^{\frac{5}{2}}\right)$, we have that $\mathbb{E}\left[\|F(\tilde{z}_0^K)\|^2\right]=\frac{2C}{n^{\frac{1}{3}}K^{\frac{1}{3}}}=\mathcal{O}\left(\frac{1}{n^{\frac{1}{3}}K^{\frac{1}{3}}}\right)$. □

C Further Convergence Guarantees

In this section, we provide theoretical guarantees for SEG-SO and IEG as well as suggest a switching stepsize rule for SEG-RR. The use of the switching stepsize rule allows us to establish for SEG-RR a $\mathcal{O}\left(\frac{1}{k}\right)$ convergence to the exact solution z_*. We highlight, also, that the proposed stepsize schedule suggests when one should switch from a constant to a decreasing stepsize regime and is to the best of our knowledge the first time used in without-replacement sampling algorithms.

C.1 Other Variants of Without-replacement Sampling

We start by showing how the proofs for SEG-RR can be modified in order to obtain convergence guarantees for two other variants of without-replacement sampling, namely the Shuffle Once (SO) sampling and the Incremental ExtraGradient (IEG).

The Shuffle Once variant samples at the first epoch of the algorithm a permutation π of the dataset and then runs SEG using one data point at each iteration of the stochastic algorithm. The data point used in the i-th iteration is π_i, namely the i-th element of the permutation π. Thus, the proofs for SEG-RR in all three regimes hold also for SEG-SO if we let $\pi^k = \pi$ for all $k \geq 0$. In this way, we are able to recover convergence guarantees for SEG-SO variant in strongly monotone, affine and monotone settings.

Regarding the Incremental ExtraGradient (IEG) variant, one can identify more easily the differences with random reshuffling in the pseudocode of Algorithm 3. Specifically, IEG does not sample any permutation of the dataset and instead regards the data samples in the order that were initially given in the dataset. Thus, the main modification in the proof of SEG-RR to get convergence guarantees for IEG is that one cannot use Lemma A.1 and instead needs to use Lemma A.3 for bounding the distance of stochastic oracles F_i from the operator F. Additionally, we observe that since the permutation $\pi = [n]$ (the initial order of the dataset) is fixed for all epochs $k \geq 0$, there is no randomness involved in the selection of the data points at each epoch and hence any term appearing in conditional expectation in the proofs of SEG-RR will be equal to the same term without the expectation in the analysis of IEG.

So far, we have explained how the proofs for SEG-RR in all three regimes can be modified to obtain convergence guarantees for SEG-SO and IEG. For illustration purposes, we provide in Sections C.1.1, C.1.2 the proof for the strongly monotone case for SEG-SO and IEG, highlighting the differences with the proof of SEG-RR. Lastly, we note that our results for the switching stepsize rule in SEG-RR from Section C.2 can be also extended to the SEG-SO and IEG algorithms.

C.1.1 SEG-SO

Corollary C.1. Suppose that the operator F is μ-strongly monotone and each F_i, $\forall i \in [n]$ is L_i-Lipschitz.

1. Then the iterates of SEG-SO with constant step size $\gamma_2 = 2\gamma_1$, $\gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$ satisfy:

$$\mathbb{E}\left[\|z_0^k - z_*\|^2\right] \leq \left(1 - \frac{\gamma_1 n\mu}{4}\right)^k \|z_0 - z_*\|^2 + \frac{96L_{max}^2}{\mu^2}\left[(25+n)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2$$

2. Let K be the total number of epochs the SEG-SO is run.
   For step size $\gamma_2 = 2\gamma_1$, $\gamma_1 = \min\left\{\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}, \frac{4\log(n^{1/2}K)}{\mu nK}\right\}$, the following holds:

$$\mathbb{E}\left[\|z_0^K - z_*\|^2\right] \quad = \quad \tilde{\mathcal{O}}\left(e^{-\frac{\mu^2 K}{L_{max}^2}} + \frac{1}{nK^2}\right)$$

Proof. Let π be the permutation that is chosen at the start of the SEG-SO algorithm. By applying Theorem 2.1 and letting the permutation $\pi^k = \pi$ for all epochs $k \geq 0$ one can observe that the algorithm run is essentially SEG-SO. Thus, the results follow immediately. □

C.1.2 IEG

Lemma C.2. Assume that each $F_i, i \in [n]$ is L_i-Lipchitz and the step size of IEG satisfy $\gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}$, $\gamma_2 \leq \frac{1}{\sqrt{n(n-1)}L_{max}}$. The iterates of the IEG algorithm satisfy the following bound

$$\frac{1}{n}\sum_{j=0}^{n-1}\left\|z_j^k - z_0^k\right\|^2 \quad \leq \quad [10n^2L^2 + 27(n-1)A]\gamma_1^2\left\|z_0^k - z_*\right\|^2 + 66n(n-1)\gamma_1^2\sigma_*^2$$

Proof. The proof of the Lemma C.2 follows exactly the proof of Lemma A.4 until inequality (31) with the only difference that the expectation of any quantity is substituted with the quantity inside the expectation. Hence, from inequality (31) we have that

$$\begin{aligned}\left\|z_i^k - z_0^k\right\|^2 \quad \leq \quad & 6\gamma_1^2L_{max}^2 i(1+2\gamma_2^2L_{max}^2)nG_k + 3\gamma_1^2(i^2 + 8\gamma_2^2 i^2 L_{max}^2)\left\|F(z_0^k)\right\|^2 \\ & +24\gamma_1^2\gamma_2^2L_{max}^2Ani\left\|z_0^k - z_*\right\|^2 + 48\gamma_1^2\gamma_2^2L_{max}^2ni\sigma_*^2 \\ & +3\gamma_1^2 i^2\left\|\frac{1}{i}\sum_{j=0}^{i-1}F_{\pi_j^k}(z_0^k) - F(z_0^k)\right\|^2\end{aligned}$$

where $G_k = \frac{1}{n}\sum\limits_{j=0}^{n-1}\left\|z_j^k - z_0^k\right\|^2$.

The only change occurs in applying Proposition A.2 to bound the last term, instead of Lemma A.3. Applying inequality (16) and Proposition A.2, we obtain

$$\begin{aligned}\left\|z_i^k - z_0^k\right\|^2 \quad \overset{(16),(A.2)}{\leq} \quad & 6\gamma_1^2L_{max}^2 i(1+2\gamma_2^2L_{max}^2)nG_k + 3\gamma_1^2(i^2 + 8\gamma_2^2 i^2L_{max}^2)\left\|F(z_0^k)\right\|^2 \\ & +3(8\gamma_2^2L_{max}^2+1)niA\gamma_1^2\left\|z_0^k - z_*\right\|^2 + 6\gamma_1^2ni(8\gamma_2^2L_{max}^2+3)\sigma_*^2\end{aligned}$$

Summing over $0 \leq i \leq n-1$ and multiplying with $\frac{1}{n}$, we get:

$$\begin{aligned}G_k \quad \leq \quad & 3\gamma_1^2L_{max}^2(1+2\gamma_2^2L_{max}^2)n(n-1)G_k + \gamma_1^2D\left\|F(z_0^k)\right\|^2 \\ & +\frac{3(8\gamma_2^2L_{max}^2+1)(n-1)A\gamma_1^2}{2}\left\|z_0^k - z_*\right\|^2 + 3\gamma_1^2n(n-1)(8\gamma_2^2L_{max}^2+3)\sigma_*^2\end{aligned}$$

where we used the fact $\frac{1}{n}\sum_{i=0}^{n-1} i = \frac{n-1}{2}$ and let also $D = \left[\frac{(1+8\gamma_2^2L_{max}^2)(n-1)(2n-1)}{2}\right]$ for brevity.

Rearranging the terms, letting $D_1 = [1 - 3n(n-1)(1+2\gamma_2^2L_{max}^2)\gamma_1^2]$ and selecting the update stepsize $\gamma_1 < \frac{1}{\sqrt{3(1+2\gamma_2^2L_{max}^2)n(n-1)}L_{max}}$, we have that

$$G_k \quad \leq \quad \gamma_1^2\frac{D}{D_1}\left\|F(z_0^k)\right\|^2 + \frac{3(8\gamma_2^2L_{max}^2+1)(n-1)A\gamma_1^2}{2D_1}\left\|z_0^k - z_*\right\|^2 + \frac{3\gamma_1^2n(n-1)(8\gamma_2^2L_{max}^2+3)\sigma_*^2}{D_1}$$

Selecting $\gamma_1 \leq \frac{1}{3\sqrt{2n(n-1)}L_{max}}, \gamma_2 \leq \frac{1}{\sqrt{n(n-1)}L_{max}}$ and using inequalities (36), (37), we get

$$G_k \quad \leq \quad 10n^2\gamma_1^2\left\|F(z_0^k)\right\|^2 + 27(n-1)A\gamma_1^2\left\|z_0^k - z_*\right\|^2 + 66n(n-1)\gamma_1^2\sigma_*^2$$

Lastly, from the Lipschitz property of F, we obtain

$$G_k \quad \leq \quad [10n^2L^2 + 27(n-1)A]\gamma_1^2\left\|z_0^k - z_*\right\|^2 + 66n(n-1)\gamma_1^2\sigma_*^2$$

□

Corollary C.3. Suppose that the operator F is μ-strongly monotone and each F_i, $\forall i \in [n]$ is L_i-Lipschitz.

1. Then the iterates of IEG with constant step size $\gamma_2 = 2\gamma_1$, $\gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+n+29}}$ satisfy:

$$\|z_0^k - z_*\|^2 \leq \left(1 - \frac{\gamma_1 n\mu}{4}\right)^k \|z_0 - z_*\|^2 + \frac{48L_{max}^2}{\mu^2}\left[6n(n-1)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2$$

2. Let K be the total number of epochs the IEG is run.
For step size $\gamma_2 = 2\gamma_1$, $\gamma_1 = \min\left\{\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+29}}, \frac{4\log(n^{1/2}K)}{\mu nK}\right\}$, the following holds:

$$\|z_0^K - z_*\|^2 \quad = \quad \tilde{\mathcal{O}}\left(e^{-\frac{K\mu^2}{L_{max}^2}} + \frac{1}{K^2}\right)$$

Proof. In IEG the data points are sampled according to the initial order in the dataset and thus $\pi = [n]$. A change to be noted in the proofs for IEG is that the algorithm does involve any stochasticity, as the permutation $\pi^k = [n]$ is fixed at each epoch $k \geq 0$. As a result, any term appearing inside expectation in the proof of SEG-RR will be deterministic in IEG and thus there is no necessity for expected values in the proofs of IEG.

By applying Theorem 2.1 and letting the permutation $\pi^k = [n]$ for all epochs $k \geq 0$ one can observe that the algorithm run is essentially IEG. The only difference with the proof of Theorem 2.1 is that in inequality (46), Lemma C.2 will be used instead of Lemma A.3 for bounding the term T_2. This will give the following upper bound

$$\|z_0^{k+1} - z_*\|^2 \quad \leq \quad \left(1 - \frac{1}{2}\gamma_1 n\mu + \frac{U}{1 - \frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu}\right)\|z_0^k - z_*\|^2 + \frac{12nL_{max}^2\gamma_1}{\mu}\left[6n(n-1)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2$$

where $C' = 2\left[10n^2L^2 + 27(n-1)A\right]\gamma_1^2 + A\gamma_2^2$. Selecting $\gamma_2 = 2\gamma_1, \gamma_1 \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+n+29}}$, we get that $\left(1 - \frac{1}{2}\gamma_1 n\mu + \frac{U}{1-\frac{\gamma_1 n\mu}{2}} + \frac{6nCL_{max}^2\gamma_1}{\mu}\right) \leq (1 - \frac{1}{4}\gamma_1 n\mu)$ and thus

$$\|z_0^{k+1} - z_*\|^2 \quad \leq \quad \left(1 - \frac{1}{4}\gamma_1 n\mu\right)\|z_0^k - z_*\|^2 + \frac{12nL_{max}^2\gamma_1}{\mu}\left[6n(n-1)\gamma_1^2 + \gamma_2^2\right]\sigma_*^2$$

Unrolling the recursion, we get

$$\|z_0^{k+1} - z_*\|^2 \quad \leq \quad \left(1 - \frac{\gamma_1 n\mu}{4}\right)^{k+1}\|z_0 - z_*\|^2 + \frac{48L_{max}^2}{\mu^2}\left[6n(n-1)\gamma_1^2 + \gamma_2^2\right] \tag{93}$$

Proof of 2nd point Substituting the stepsize $\gamma_1 = \min\left\{\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+29}}, \frac{4\log(n^{1/2}K)}{\mu nK}\right\} \leq \frac{4\log(n^{1/2}K)}{\mu nK}$ into (93), we have that

$$\begin{aligned}\|z_0^K - z_*\|^2 \quad &\leq \quad \left(1 - \frac{1}{4}\gamma_1 n\mu\right)^K \|z_0 - z_*\|^2 + \frac{48L_{max}^2}{\mu^2}\left[6n(n-1) + 4\right]\frac{16\log^2(n^{1/2}K)}{\mu^2 n^2 K^2}\sigma_*^2 \\ &\leq \quad \left(1 - \frac{1}{4}\gamma_1 n\mu\right)^K \|z_0 - z_*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{K^2}\right)\end{aligned}$$

We now consider the following cases:

Case 1: $\frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+29}} \leq \frac{\log(n^{1/2}K)}{\mu nK}$ In this case we have that $\gamma_1 = \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$, which implies that the RHS of (57) is bounded by

$$\begin{aligned}\|z_0^K - z_*\|^2 \quad &\leq \quad e^{-\frac{\gamma_1 nK\mu}{4}}\|z_0 - z_*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{K^2}\right) \\ &\leq \quad e^{-\frac{nK\mu^2}{40L_{max}^2\sqrt{10n^2+2n+29}}}\|z_0 - z_*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{K^2}\right) \\ &\leq \quad e^{-\frac{K\mu^2}{40\sqrt{12}L_{max}^2}}\|z_0 - z_*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{K^2}\right)\end{aligned} \tag{94}$$

Case 2: $\frac{4\log(n^{1/2}K)}{\mu n K} \leq \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+29}}$ In this case we have that $\gamma_1 = \frac{4\log(n^{1/2}K)}{\mu n K}$, which implies that the RHS of (57) is bounded by

$$\begin{aligned} \|z_0^K - z_*\|^2 &\leq e^{-\frac{\gamma_1 n K \mu}{4}}\|z_0 - z_*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{K^2}\right) \\ &\leq \frac{1}{nK^2}\|z_0 - z_*\|^2 + \tilde{\mathcal{O}}\left(\frac{1}{K^2}\right) \end{aligned} \tag{95}$$

Taking the maximum of the right-hand side of (94) and (95) and using the inequality $\max\{a,b\} \leq a + b$, we obtain the desired result which holds for both cases:

$$\|z_0^K - z_*\|^2 \leq e^{-\frac{K\mu^2}{40\sqrt{12}L_{max}^2}\|z_0-z_*\|^2+\frac{1}{K^2}\|z_0-z_*\|^2+\tilde{\mathcal{O}}(\frac{1}{K^2})}$$

Suppressing constant and logarithmic terms, we get the final result

$$\|z_0^K - z_*\|^2 \;=\; \tilde{\mathcal{O}}\left(e^{-\frac{K\mu^2}{L_{max}^2}} + \frac{1}{K^2}\right)$$

□

C.2 SEG-RR with Switching Stepsize Rule

We, next, provide theorems for the use of a switching stepsize rule in the strongly monotone and affine case that allows us to establish convergence to the exact solution z_*. The stepsize rule indicates the use of a constant stepsize at the start of the algorithm in order to converge to a neighborhood around the solution z_* and then switch to a decreasing one with the goal of reducing the neighborhood and converging to the exact solution.

Theorem C.4. Suppose that the operator F is μ-strongly monotone, each F_i, $\forall i \in [n]$ is L_i-Lipschitz and SEG-RR is run with step size $\gamma_{2,k} = 2\gamma_{1,k}$,

$$\gamma_{1,k} = \begin{cases} \gamma_{1,max}, \text{ for } k < k^* = \left\lceil \dfrac{64}{\mu^2\gamma_{1,max}^2} \right\rceil \\ \dfrac{4(2k+1)}{\mu(k+1)^2}, \quad \text{for } k \geq k^* \end{cases}$$

Then, we have that the iterates of SEG-RR satisfy

$$\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2\right] \;\leq\; \frac{(k^*)^2 e^{-\frac{\mu n \gamma_{1,max}}{4}}}{(K+1)^2}\|z_0 - z_*\|^2 + \frac{96L_{max}^2(29+n)\sigma_*^2}{\mu^2(K+1)^2}\left(\gamma_{1,max}^2 {k^*}^2 + \frac{128}{\mu^2}K\right)$$

where $\gamma_{1,max} = \frac{\mu}{10L_{max}^2\sqrt{10n^2+2n+54}}$.

Proof. Let $\gamma_{1,k}, \gamma_{2,k}$ be the step size of SEG-RR algorithm in the k-th epoch and fix $\gamma_{2,k} = 2\gamma_{1,k}$. Let also $k^* \in \mathbb{Z}_*$ be an epoch at which the stepsize scheme uses the decreasing stepsize $\gamma_{1,k} = \frac{4(2k+1)}{\mu(k+1)^2}$ and satisfies $\gamma_{1,k^*} \leq \gamma_{1,max}$. Observe that $\forall k \geq k^* : \gamma_{1,k} \leq \gamma_{1,max}$ and thus inequality (52) holds.

Substituting $\gamma_1 = \gamma_{1,k}$ and $\gamma_{2,k} = 2\gamma_{1,k}$ into (52), we get:

$$\begin{aligned} \mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] &\leq \left(1 - \frac{1}{4}\gamma_{1,k}n\mu\right)\|z_0^k - z_*\|^2 + \frac{24nL_{max}^2(29+n)}{\mu}\gamma_{1,k}^3\sigma_*^2 \\ &\leq \left(1 - \frac{1}{4}\gamma_{1,k}n\mu\right)\|z_0^k - z_*\|^2 + \frac{24nL_{max}^2(29+n)}{\mu}\frac{4^3(2k+1)^3}{\mu^3(k+1)^6}\sigma_*^2 \end{aligned} \tag{96}$$

We, then, multiply both sides of (96) by $(k+1)^2$ and obtain

$$(k+1)^2\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \leq k^2\mathbb{E}\left[\|z_0^{k^*} - z_*\|^2\right] + \frac{24nL_{max}^2(29+n)}{\mu^4}\frac{4^3(2k+1)^3}{(k+1)^4}\sigma_*^2$$

Using the inequality $\frac{(2k+1)^3}{(k+1)^4} \leq 8$, we have that:

$$(k+1)^2 \mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \leq k^2 \|z_0^k - z_*\|^2 + \frac{48 \cdot 4^4 L_{max}^2 (29+n)}{\mu^4} \sigma_*^2$$

Rearranging the terms and summing for $k = k^*, ..., K$ we are able to get the telescopic cancellation

$$\sum_{k=k^*}^{K} \left[(k+1)^2 \mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] - k^2 \|z_0^k - z_*\|^2\right] \quad \leq \quad \sum_{k=k^*}^{K} \frac{48 \cdot 4^4 L_{max}^2 (29+n) \sigma_*^2}{\mu^4}$$

Thus, we have that:

$$\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad \frac{(k^*)^2}{(K+1)^2} \mathbb{E}\left[\|z_0^{k^*} - z_*\|^2 \middle| \mathcal{F}^k\right] + \frac{48 \cdot 4^4 L_{max}^2 (29+n)}{\mu^4 (K+1)^2} (K - k^*) \sigma_*^2$$

Taking expectation on both sides and using the tower property, we get

$$\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2\right] \quad \leq \quad \frac{(k^*)^2}{(K+1)^2} \mathbb{E}\left[\|z_0^{k^*} - z_*\|^2\right] + \frac{48 \cdot 4^4 L_{max}^2 (29+n)}{\mu^4 (K+1)^2} (K - k^*) \sigma_*^2 \tag{97}$$

For $k \leq k^*$ we have that (54) holds and thus combining it with (97) results to

$$\begin{aligned} \mathbb{E}\left[\|z_0^{K+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad & \frac{(k^*)^2}{(K+1)^2} \left(1 - \frac{\mu n \gamma_{1,max}}{4}\right)^{k^*} \|z_0 - z_*\|^2 \\ & + \frac{96 L_{max}^2 (29+n) \sigma_*^2}{\mu^2 (K+1)^2} \left(\gamma_{1,max}^2 {k^*}^2 + \frac{128}{\mu^2} (K - k^*)\right) \end{aligned} \tag{98}$$

Using the inequality $(1-x)^x \leq e^{-x}$, we get:

$$\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2\right] \quad \leq \quad \frac{(k^*)^2 e^{-\frac{\mu n \gamma_{1,max}}{4}}}{(K+1)^2} \|z_0 - z_*\|^2 + \frac{96 L_{max}^2 (29+n) \sigma_*^2}{\mu^2 (K+1)^2} \left(\gamma_{1,max}^2 {k^*}^2 + \frac{128}{\mu^2} K\right) \tag{99}$$

In the above, we choose k^* so that it minimizes the second term in (99) and thus $k^* = \left\lceil \frac{64}{\mu^2 \gamma_{1,max}^2} \right\rceil$. □

Next, we provide convergence guarantees for a switching stepsize rule in the affine case.

Theorem C.5. Suppose that each F_i, $\forall i \in [n]$ is monotone, affine and L_i−Lipschitz. If SEG-RR is run with step size $\gamma_{2,k} = 4\gamma_{1,k}$,

$$\gamma_{1,k} = \begin{cases} \gamma_{1,max}, & \text{for } k < k^* = \left\lceil \frac{16}{\lambda_{min}^+(Q)^2 \gamma_{1,max}^2} \right\rceil \\ \frac{2(2k+1)}{\lambda_{min}^+(Q)(k+1)^2}, & \text{for } k \geq k^* \end{cases}$$

then we have that the iterates of SEG-RR satisfy

$$\begin{aligned} \mathbb{E}\left[\|z_0^{K+1} - z_*\|^2\right] \quad \leq \quad & \frac{(k^*)^2}{(K+1)^2} e^{-\frac{1}{2}\gamma_1 n k^* \lambda_{min}^+(Q)} \mathbb{E}\left[\|z_0 - z_*\|^2\right] \\ & + \frac{8\left(24n - 23 + \frac{1}{n}\right) L_{max}}{\lambda_{min}^+(Q)^2 (K+1)^2} \sigma_*^2 \left[\gamma_{1,max}^2 {k^*}^2 + \frac{32(K - k^*)}{\lambda_{min}^2(Q)}\right] \end{aligned}$$

where $\gamma_{1,max} = \frac{\lambda_{min}^+(Q)}{2\sqrt{120 n L_{max}^2}}$.

Proof. Let $\gamma_{1,k}, \gamma_{2,k}$ be the step size of SEG-RR algorithm in the k-th epoch and fix $\gamma_{2,k} = 4\gamma_{1,k}$. Let also $k^* \in \mathbb{Z}_*$ be an epoch at which the stepsize scheme uses the decreasing stepsize $\gamma_{1,k} = \frac{2(2k+1)}{\lambda^+_{min}(Q)(k+1)^2}$. Observe that $\forall k \geq k^* : \gamma_{1,k} \leq \gamma_{1,max}$ and thus inequality (69) holds.
Substituting $\gamma_1 = \gamma_{1,k}$ and $\gamma_2 = 4\gamma_{1,k}$ in (69) we get:

$$\begin{aligned}
\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \left[1 - \frac{2k+1}{(k+1)^2}\right]\|z_0^k - z_*\|^2 \\
&\quad + \frac{4\left(24n^2 - 23n + 1\right)L_{max}\sigma_*^2}{\lambda^+_{min}(Q)}\left(\frac{2(2k+1)}{\lambda^+_{min}(Q)(k+1)^2}\right)^3
\end{aligned}$$

Multiplying both sides with $(k+1)^2$ results to

$$(k+1)^2\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \leq k^2\mathbb{E}\left[\|z_0^{k^*} - z_*\|^2 \middle| \mathcal{F}^k\right] + \frac{32\left(24n^2 - 23n + 1\right)L_{max}\sigma_*^2}{\lambda^+_{min}(Q)^4}\frac{(2k+1)^3}{(k+1)^4}$$

Using the inequality $\frac{(2k+1)^3}{(k+1)^4} \leq \frac{2^3(k+1)^3}{(k+1)^4} \leq 8$ we have that:

$$(k+1)^2\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \leq k^2\|z_0^k - z_*\|^2 + \frac{256\left(24n^2 - 23n + 1\right)L_{max}\sigma_*^2}{\lambda^+_{min}(Q)^4}$$

Rearranging the terms and summing for $k = k^*, ..., K$, we are able to get the telescopic cancellation

$$\sum_{k=k^*}^{K}\left[(k+1)^2\mathbb{E}\left[\|z_0^{k+1} - z_*\|^2 \middle| \mathcal{F}^k\right] - k^2\|z_0^k - z_*\|^2\right] \quad \leq \quad \sum_{k=k^*}^{K}\frac{256\left(24n^2 - 23n + 1\right)L_{max}\sigma_*^2}{\lambda^+_{min}(Q)^4}$$

Thus, we get that:

$$\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad \leq \quad \frac{(k^*)^2}{(K+1)^2}\mathbb{E}\left[\|z_0^{k^*} - z_*\|^2 \middle| \mathcal{F}^k\right] + \frac{256\left(24n^2 - 23n + 1\right)L_{max}\sigma_*^2}{\lambda^+_{min}(Q)^4(K+1)^2}$$

Taking expectation on both sides and using the tower property, we obtain:

$$\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2\right] \quad \leq \quad \frac{(k^*)^2}{(K+1)^2}\mathbb{E}\left[\|z_0^{k^*} - z_*\|^2\right] + \frac{256\left(24n^2 - 23n + 1\right)L_{max}\sigma_*^2}{\lambda^+_{min}(Q)^4(K+1)^2}(K - k^*)\sigma_*^2 \tag{100}$$

For $k \leq k^*$ we have that (72) holds with $\gamma_1 = \gamma_{1,max}$

$$\mathbb{E}\left[\|z_0^{k^*+1} - z_*\|^2\right] \quad \leq \quad \left[1 - \frac{1}{2}\gamma_{1,max}n\lambda^+_{min}(Q)\right]^{k^*}\|z_0 - z_*\|^2 + \frac{8\left(24n - 23 + \frac{1}{n}\right)L_{max}}{\lambda^2_{min}(Q)}\gamma^2_{1,max}\sigma^2_*$$

and thus combining it with (100) we get:

$$\begin{aligned}
\mathbb{E}\left[\|z_0^{K+1} - z_*\|^2 \middle| \mathcal{F}^k\right] \quad &\leq \quad \frac{(k^*)^2}{(K+1)^2}\left(1 - \frac{1}{2}\gamma_1 n\lambda^+_{min}(Q)\right)^{k^*}\mathbb{E}\left[\|z_0 - z_*\|^2\right] \\
&\quad + \frac{8\left(24n - 23 + \frac{1}{n}\right)L_{max}}{\lambda^2_{min}(Q)(K+1)^2}\sigma^2_*\left[\gamma^2_{1,max}{k^*}^2 + \frac{32(K-k^*)}{\lambda^2_{min}(Q)}\right] \\
&\overset{(19)}{\leq} \quad \frac{(k^*)^2}{(K+1)^2}e^{-\frac{1}{2}\gamma_1 nk^*\lambda^+_{min}(Q)}\mathbb{E}\left[\|z_0 - z_*\|^2\right] \\
&\quad + \frac{8\left(24n - 23 + \frac{1}{n}\right)L_{max}}{\lambda^2_{min}(Q)(K+1)^2}\sigma^2_*\left[\gamma^2_{1,max}{k^*}^2 + \frac{32(K-k^*)}{\lambda^2_{min}(Q)}\right]
\end{aligned} \tag{101}$$

Lastly, we choose k^*, so that it minimizes the second term in (101) and thus $k^* = \left\lceil\frac{16}{\lambda^2_{min}(Q)\gamma^2_{1,max}}\right\rceil$. □

D On Experiments

In Appendix D.1, we provide more details on the experiments discussed in the main paper. In Appendix D.2, we run more experiments to evaluate the performance of SEG-RR. As stated in the main paper, the code for reproducing our experimental results is available at `https://github.com/emmanouilidisk/Stochastic-ExtraGradient-with-RR`.

D.1 Experimental Details

We first describe our experimental setup. In the strongly monotone setting, we consider the following quadratic problem:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}x^\top A_i x + x^\top B_i y - \frac{1}{2}y^\top C_i y + a_i^\top x - c_i^\top y$$

We sample the matrices A_i by first sampling an orthogonal matrix P and then sampling a diagonal matrix D_i with elements in the diagonal uniformly sampled from the interval $[\mu, L]$. Here, the parameters μ, L correspond to the strong monotonicity parameter and the Lipschitz parameter of the problem. We acquire the matrices A_i, as the product $A_i = PD_iP^T$. We sample the matrices B_i, C_i similarly to sampling the matrices A_i with the only difference that the elements of D_i lie in the interval $[0, 0.1]$ and $[\mu, L]$ respectively. The vectors a_i, c_i are sampled from the normal distribution $\mathcal{N}(0, I)$. In all experiments, we use $n = 100, d = 100$, while we specify the values of μ, L in each experiment independently as they differ.

In the bilinear regime, we focus on the following two-player zero-sum game:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}x^\top B_i y + a_i^\top x - c_i^\top y$$

We let the matrices B_i be $B_i = PD_iP^T$, where P is an orthogonal matrix and D_i a diagonal matrix with elements in the diagonal selected uniformly at random from $[\lambda^+_{min}(Q), L_{max}]$. We specify that the parameters $\lambda^+_{min}(Q), L_{max}$ correspond to the parameters $\lambda^+_{min}(Q), L_{max}$ of Theorem 2.2. Regarding the vectors a_i, c_i, they are sampled from the normal distribution $\mathcal{N}(0, I)$. In all experiments, we use $n = 100$ and let $d = 1$ in the two-dimensional experiments; otherwise, $d = 100$. We specify individually for each experiment the parameters $\lambda^+_{min}(Q), L_{max}$ that have been used.

D.2 Additional Experiments

In this part, we provide additional experiments to the ones presented in the main paper.

SC - SC Problems. We initially focus on Strongly Convex - Strongly Concave (SC - SC) minimax problems and compare the different without-replacement sampling variants of SEG, namely SEG-RR, SEG-SO and IEG, with S-SEG (denoted as *SEG* in the plots).

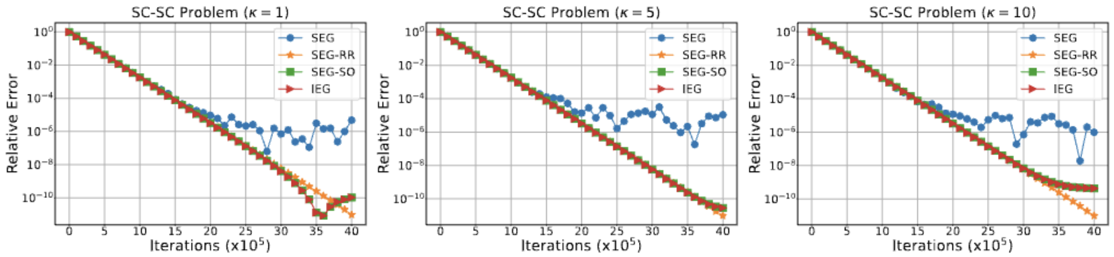

Figure 6: SC-SC problem. SEG-RR, SEG-SO, SEG-IG and SEG run with step size as in Theorem 2.1 for problem with $\mu = 1, n = 100$ and different condition numbers. SEG-RR achieves smaller relative error in comparison to the other without-replacement sampling methods.

Since SEG-RR seems to achieve at least as small (if not smaller) error than the other without-replacement variants it makes sense to use SEG-RR in practice. In this way, we will focus for the rest of this section on experiments comparing SEG-RR with SEG. We start by exploiting the behaviour of with and without-replacement sampling for the step size suggested by the analysis of SEG in Gorbunov et al. [2022a]. We observe that even for the theoretical step size SEG-RR performs better than SEG in terms of relative error.

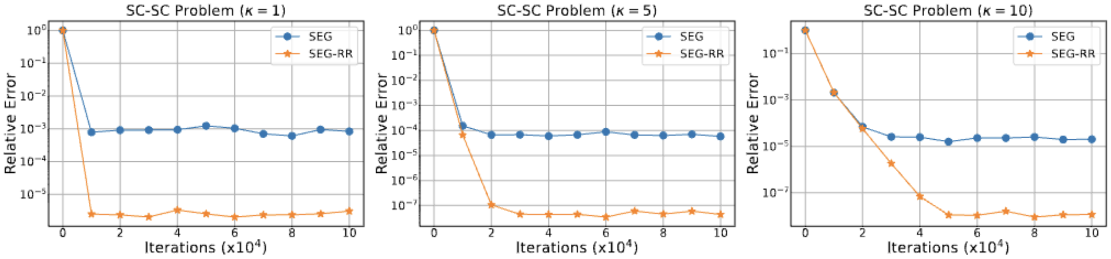

Figure 7: SC-SC problem. SEG-RR vs SEG run with step size as in Gorbunov et al. [2022a] for problem with $\mu = 1, n = 100$ and different condition numbers.

We, next, compare SEG-RR with SEG. We conduct experiments for problems with different condition numbers $\kappa = \{1, 5, 10, 100\}$ and different step size $\gamma_1 = \{\frac{1}{10L_{max}}, \frac{1}{100L_{max}}, \frac{1}{1000L_{max}}\}$. In this vein, we fix $\mu = 1$ and let the Lipschitz parameter vary as the condition number $\kappa = \frac{L}{\mu}$ changes.

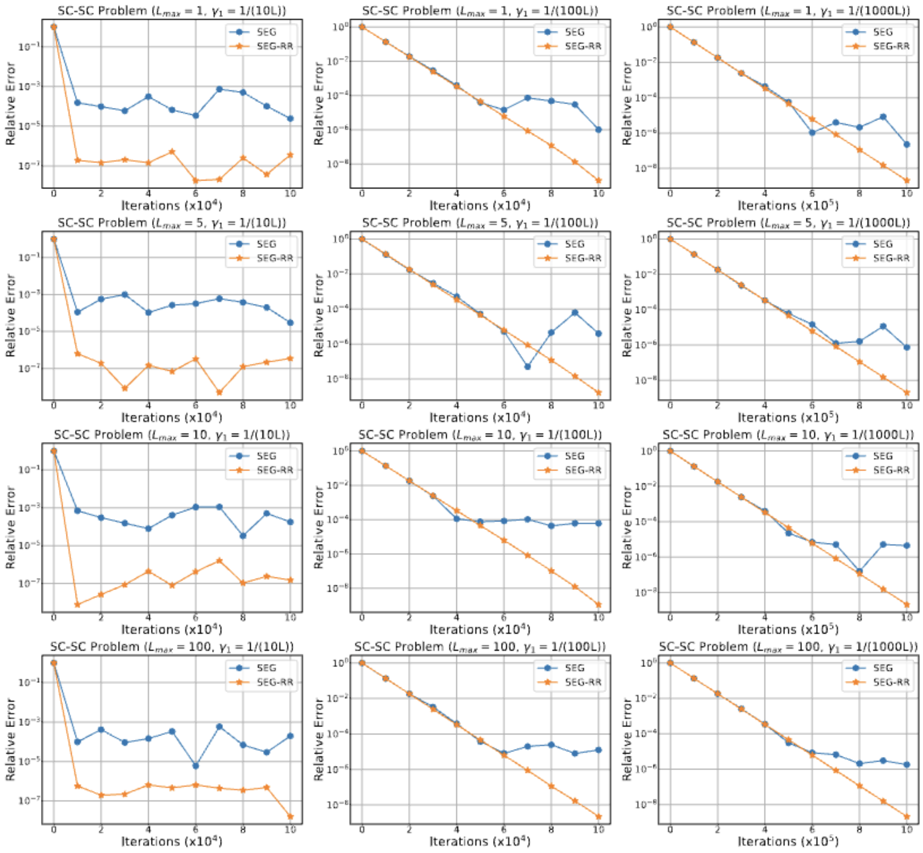

Figure 8: SC - SC Problems: Comparison of SEG-RR and SEG with different step size. Each row corresponds to a problem with a different condition number $\kappa = \{1, 5, 10, 100\}$, while each column corresponds to a specific stepsize $\gamma_1 = \{\frac{1}{10L}, \frac{1}{100L}, \frac{1}{1000L}\}$. In all problems $\mu = 1, n = 100$.

Bilinear Games. We, first, provide experiments comparing SEG-RR, SEG-SO, IEG with SEG. We use as step size in all algorithms the step size suggested in Theorem 2.2. We fix $\lambda_{min}^{+}(Q) = 1$ and let the Lipschitz constant of the problem vary.

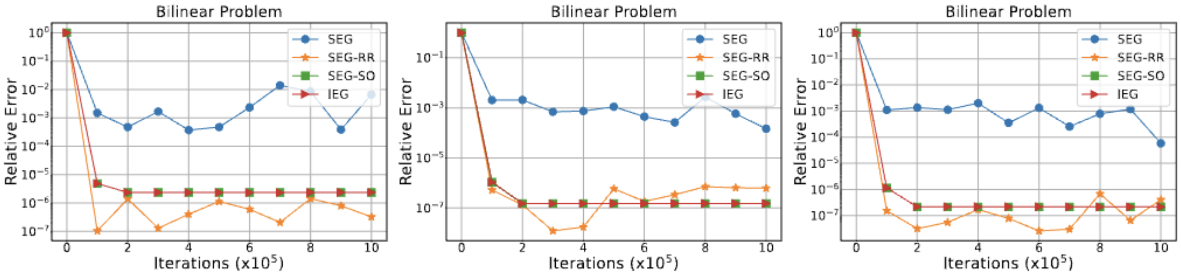

Figure 9: Bilinear games. The without-replacement variants of SEG achieve better performance in terms of relative error in comparison to S-SEG. The problem parameters $\lambda_{min}^{+}(Q) = 1, n = 100$ for different $L_{max} = \{1, 5, 10\}$ in the plots from left to right accordingly and with step size the ones in Theorem 2.2.

It is obvious that the without-replacement sampling variants of SEG converge with with smaller relative error than the uniform with-replacement variant for the same number of iterations/epochs.

We, next, provide experiments for SEG-RR and SEG for problems with different Lipschitz parameters $L_{max} = \{1, 5, 10\}$ using the theoretical step size $\gamma_1 = \frac{0.1}{(t+19)^{r_\eta}}, \gamma_2 = \frac{1}{(t+19)^{r_\gamma}}$ where $r_\gamma = 0, r_\eta = 0.7$ suggested in the analysis of SEG for bilinear games in Hsieh et al. [2020].

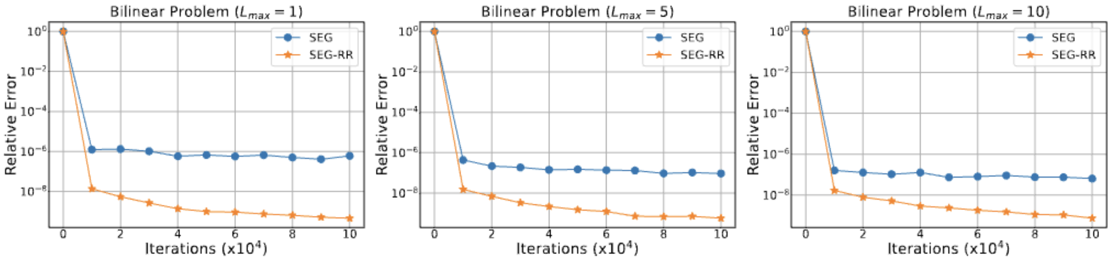

Figure 10: Bilinear games with $\lambda_{min}^{+}(Q) = 1, n = 100$ for different L_{max} and with step sizes as in Theorem 2.2.

It is obvious that for the stepsizes suggested by theory SEG-RR achieves a smaller relative error for the same number of epochs/iterations in comparison to SEG.

We, lastly, conduct experiments for step sizes larger than the theoretical ones and for a number of different problem instances to capture the performance of SEG-RR and SEG in a broad range of step sizes and problem setups. We run experiments for problems with different $L = \{1, 5, 10\}$ and for step sizes $\gamma_2 = 4\gamma_1$ with $\gamma_1 = \{\frac{1}{10L_{max}}, \frac{1}{100L_{max}}\}$.

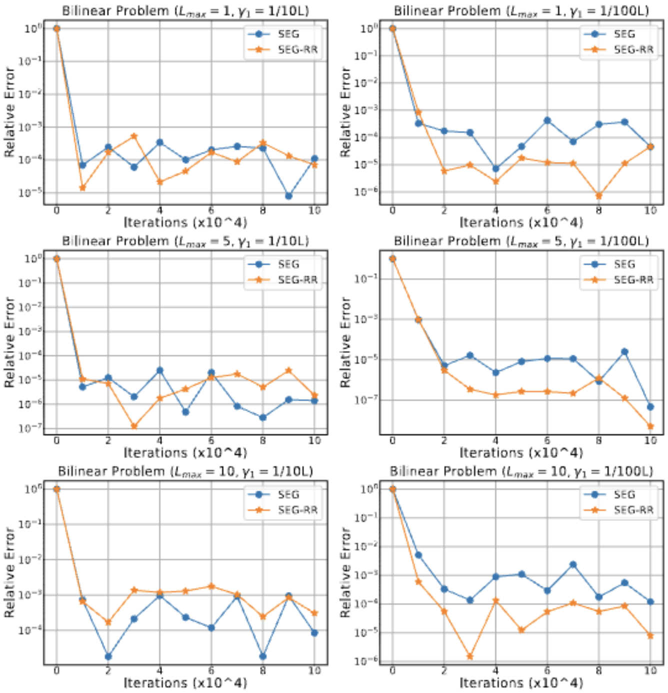

Figure 11: Bilinear games for different $L_{max} = \{1, 5, 10\}$ and with step sizes $\gamma_1 = \{\frac{1}{10L}, \frac{1}{100L}\}$. In all problems $\lambda^+_{min}(Q) = 1, n = 100$.

In the above plots, it is obvious that in most cases SEG-RR achieves at least as good (if not smaller) relative error than S-SEG, which advocates for the use of random reshuffling in practice.