Stochastic Gradient Descent-Ascent: Unified Theory and New Efficient Methods

Aleksandr Beznosikov*
Innopolis University,
HSE University,
Yandex

Eduard Gorbunov*
MBZUAI

Hugo Berard*
Mila and DIRO,
Université de Montréal

Nicolas Loizou
AMS and MINDS,
Johns Hopkins University

Abstract

Stochastic Gradient Descent-Ascent (SGDA) is one of the most prominent algorithms for solving min-max optimization and variational inequalities problems (VIP) appearing in various machine learning tasks. The success of the method led to several advanced extensions of the classical SGDA, including variants with arbitrary sampling, variance reduction, coordinate randomization, and distributed variants with compression, which were extensively studied in the literature, especially during the last few years. In this paper, we propose a unified convergence analysis that covers a large variety of stochastic gradient descent-ascent methods, which so far have required different intuitions, have different applications and have been developed separately in various communities. A key to our unified framework is a parametric assumption on the stochastic estimates. Via our general theoretical framework, we either recover the sharpest known rates for the known special cases or tighten them. Moreover, to illustrate the flexibility of our approach, we develop several new variants of SGDA such as a new variance-reduced method (L-SVRGDA), new distributed methods with compression (QSGDA, DIANA-SGDA, VR-DIANA-SGDA), and a new method with coordinate randomization (SEGA-SGDA). Although variants of the new methods are known for solving minimization problems, they were never considered or analyzed for solving min-max problems and VIPs. We also demonstrate the most important properties of the new methods through extensive numerical experiments.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.
*Equal contribution.

1 INTRODUCTION

Min-max optimization and, more generally, variational inequality problems (VIPs) appear in a wide range of research areas, including but not limited to statistics (Bach, 2019), online learning (Cesa-Bianchi and Lugosi, 2006), game theory (Morgenstern and Von Neumann, 1953), and machine learning (Goodfellow et al., 2014). Motivated by applications in these areas, in this paper, we focus on solving the following regularized VIP: Find $x^* \in \mathbb{R}^d$ such that

$$\langle F(x^*), x - x^* \rangle + R(x) - R(x^*) \geq 0 \quad \forall x \in \mathbb{R}^d, \quad (1)$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is some operator and $R : \mathbb{R}^d \to \mathbb{R}$ is a regularization term (a proper lower semicontinuous convex function), which is assumed to have a simple structure. This problem is quite general and covers a wide range of possible problem formulations. For example, when operator $F(x)$ is the gradient of a convex function f, then problem (1) is equivalent to the composite minimization problem (Beck, 2017), i.e., minimization of $f(x) + R(x)$. Problem (1) is also a more abstract formulation of the min-max problem

$$\min_{x_1 \in Q_1} \max_{x_2 \in Q_2} f(x_1, x_2), \quad (2)$$

with convex-concave continuously differentiable f. In that case, first-order optimality conditions imply that (2) is equivalent to (1) with $x = (x_1^\top, x_2^\top)^\top$, $F(x) = (\nabla_{x_1} f(x_1, x_2)^\top, -\nabla_{x_2} f(x_1, x_2)^\top)^\top$, and $R(x) = \delta_{Q_1}(x_1) + \delta_{Q_2}(x_2)$, where $\delta_Q(\cdot)$ is an indicator function of the set Q (Alacaoglu and Malitsky, 2021). In addition, to formulate the constraints, regularization R allows us to enforce some properties to the solution x^*, e.g., sparsity (Candes et al., 2008; Beck, 2017).

More precisely, we are interested in the situations when operator F is accessible through the calls of unbiased stochastic oracle. This is natural when F has an expectation form $F(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F_\xi(x)]$ or a finite-sum form $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$. In the context of machine learning, $\mathcal{D}$ corresponds to some unknown distribution on the data, n corresponds to the number of samples, and F_ξ, F_i denote vector fields corresponding to the samples ξ, and i, respectively (Gidel et al., 2019; Loizou et al., 2021).

One of the most popular methods for solving (1) is Stochastic Gradient Descent-Ascent[1] (SGDA) (Dem'yanov and Pevnyi, 1972; Nemirovski et al., 2009). However, besides its rich history, SGDA only recently was analyzed without using strong assumptions on the noise (Loizou et al., 2021) such as uniformly bounded variance. In the last few years, several powerful algorithmic techniques like variance reduction (Palaniappan and Bach, 2016; Yang et al., 2020) and coordinate-wise randomization (Sadiev et al., 2021), were also combined with SGDA resulting in better algorithms. However, these methods were analyzed under different assumptions, using different analysis approaches, and required different intuitions. Moreover, to the best of our knowledge, fruitful directions such as communication compression for distributed versions of SGDA or linearly converging variants of coordinate-wise methods for regularized VIPs were never considered in the literature before.

All of these facts motivate the importance and necessity of a novel general analysis of SGDA unifying several special cases and providing the ability to design and analyze new SGDA-like methods filling existing gaps in the theoretical understanding of the method.

In this work, we develop such unified analysis.

1.1 Technical Preliminaries

Throughout the paper, we assume that (1) has at least one solution and operator F is *μ-quasi-strongly monotone* and *ℓ-star-cocoercive*: there exist constants $\mu \geq 0$ and $\ell > 0$ such that for all $x \in \mathbb{R}^d$

$$\langle F(x) - F(x^*), x - x^* \rangle \geq \mu \|x - x^*\|^2, \tag{3}$$

$$\|F(x) - F(x^*)\|^2 \leq \ell \langle F(x) - F(x^*), x - x^* \rangle, \tag{4}$$

where $x^* = \mathrm{proj}_{X^*}(x) := \arg\min_{y \in X^*} \|y - x\|$ is the projection of x on the solution set X^* of (1). If $\mu = 0$, inequality (3) is known as variational stability condition Hsieh et al. (2020), which is weaker than standard monotonicity: $\langle F(x) - F(y), x - y \rangle \geq 0$ for all $x, y \in \mathbb{R}^d$. It is worth mentioning that there exist examples of non-monotone operators satisfying (3) with $\mu > 0$ (Loizou et al., 2021). Condition (4) is a relaxation of standard cocoercivity $\|F(x) - F(y)\|^2 \leq \ell \langle F(x) - F(y), x - y \rangle$. At this point, let us highlight that it is possible for an operator F to satisfy (4) and not be Lipschitz continuous (Loizou et al., 2021). This emphasizes the wider applicability of the *ℓ-star-cocoercivity* compared to ℓ-cocoercivity. We emphasize that in our convergence analysis, we do not assume ℓ-cocoercivity nor L-Lipschitzness of F.

We consider SGDA for solving (1) in its general form:

$$x^{k+1} = \mathrm{prox}_{\gamma_k R}(x^k - \gamma_k g^k), \tag{5}$$

[1] This name is usually used in the min-max setup. Although we consider a more general problem formulation, we keep the name SGDA to highlight the connection with min-max problems.

where g^k is an unbiased estimator of $F(x^k)$, $\gamma_k > 0$ is a stepsize at iteration k, and $\mathrm{prox}_{\gamma R}(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ R(y) + \|y-x\|^2/2\gamma \right\}$ is a proximal operator defined for any $\gamma > 0$ and $x \in \mathbb{R}^d$. While g^k gives an information about operator F at step k, proximal operator is needed to take into account regularization term R. We assume that function R is such that $\mathrm{prox}_{\gamma R}(x)$ can be easily computed for all $x \in \mathbb{R}^d$. This is a standard assumption satisfied for many practically interesting regularizers (Beck, 2017). By default we assume that $\gamma_k \equiv \gamma > 0$ for all $k \geq 0$.

1.2 Our Contributions

- **Unified analysis of SGDA.** We propose a general assumption on the stochastic estimates and the problem (1) (Assumption 2.1) and show that several variants of SGDA (5) satisfy this assumption. In particular, through our approach, we cover SGDA with arbitrary sampling (Loizou et al., 2021), variance reduction, coordinate randomization, and compressed communications. Under Assumption 2.1 we derive general convergence results for quasi-strongly monotone (Theorem 2.2), monotone star-cocoercive (Theorem 2.5) and cocoercive problems (Theorem 2.6).

- **Extensions of known methods and analysis.** As a by-product of the generality of our theoretical framework, we derive new results for the proximal extensions of several known methods such as *proximal* SGDA-AS (Loizou et al., 2021) and *proximal* SGDA with coordinate randomization (Sadiev et al., 2021). Moreover, we close some gaps on the convergence of known methods, e.g., we derive the first convergence guarantees in the monotone case for SGDA-AS (Loizou et al., 2021) and SAGA-SGDA (Palaniappan and Bach, 2016) and we obtain the first result on the convergence of SAGA-SGDA for (averaged star-)cocoercive operators.

- **Sharp rates for known special cases.** For the known methods fitting our framework our general theorems either recover the best rates known for these methods (SGDA-AS) or tighten them (SGDA-SAGA, Coordinate SGDA).

- **New methods.** The flexibility of our approach allows us to develop and analyze several new variants of SGDA. Guided by algorithmic advances for solving minimization problems we propose a new variance-reduced method (L-SVRGDA), new distributed methods with compression (QSGDA, DIANA-SGDA, VR-DIANA-SGDA), and a new method with coordinate randomization (SEGA-SGDA). We show that the proposed new methods fit our theoretical framework and, using our general theorems, we obtain tight convergence guarantees for them. Although the analogs of these methods are known for solving minimization problems (Hofmann et al., 2015; Kovalev et al., 2020; Alistarh et al., 2017;

Mishchenko et al., 2019; Horváth et al., 2019; Hanzely et al., 2018), they were never considered for solving min-max and variational inequality problems. Therefore, by proposing and analyzing these new methods we close several gaps in the literature on SGDA. For example, VR-DIANA-SGDA is the first SGDA-type *linearly converging* distributed stochastic method with compression and SEGA-SGDA is the first *linearly converging* coordinate method for solving *regularized* VIPs.

◇ **Numerical evaluation.** In numerical experiments, we illustrate the most important properties of the new methods. The results corroborate our theoretical findings.

Throughout the paper, we provide necessary comparisons with closely related work. Additional works relevant to our paper are discussed in Appendix A.

2 UNIFIED ANALYSIS OF SGDA

Key assumption. We start by introducing the next parametric assumption – a central part of our approach.

Assumption 2.1. We assume that for all $k \geq 0$ the estimator g^k from (5) is unbiased: $\mathbb{E}_k\left[g^k\right] = F(x^k)$, where $\mathbb{E}_k[\cdot]$ denotes the expectation w.r.t. the randomness at iteration k. Next, we assume that there exist non-negative constants $A, B, C, D_1, D_1 \geq 0$, $\rho \in (0,1]$ and a sequence of (possibly random) non-negative variables $\{\sigma_k\}_{k\geq 0}$ such that for all $k \geq 0$

$$\begin{aligned} \mathbb{E}_k\left[\|g^k - g^{*,k}\|^2\right] &\leq 2A\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle \\ &\quad + B\sigma_k^2 + D_1, \qquad (6) \\ \mathbb{E}_k\left[\sigma_{k+1}^2\right] &\leq 2C\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle \\ &\quad + (1-\rho)\sigma_k^2 + D_2, \qquad (7) \end{aligned}$$

where $x^{*,k} = \mathrm{proj}_{X^*}(x^k)$ and $g^{*,k} = F(x^{*,k})$.

While unbiasedness of g^k is a standard assumption, inequalities (6)-(7) are new and require clarifications. For simplicity, assume that $\sigma_k^2 \equiv 0$, $F(x^*) = 0$ for all $x^* \in X^*$, and focus on (6). In this case, (6) gives an upper bound for the second moment of the stochastic estimate g^k. For example, such a bound follows from expected cocoercivity assumption (Loizou et al., 2021), where A denotes some expected/averaged (star-)cocoercivity constant and D_1 stands for the variance at the solution (see also Section 3). When F is not necessarily zero on X^*, the shift $g^{*,k}$ helps to take this fact into account. Finally, the sequence $\{\sigma_k^2\}_{k\geq 0}$ is typically needed to capture the variance reduction process, parameter B is typically some numerical constant, C is another constant related to (star-)cocoercivity[2], and D_2 is the remaining noise that is not handled by the variance reduction process. As we show in the next sections, inequalities (6)-(7) hold for various SGDA-type methods.

We point out that Assumption 2.1 is inspired by similar assumptions appeared in Gorbunov et al. (2020a, 2022a). However, the difference between our assumption and the ones appeared in these papers is significant: Gorbunov et al. (2020a) focuses only on solving minimization problems and as a result, their assumption includes a much simpler quantity (function suboptimality), instead of the $\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle$, in the right-hand sides of (6)-(7). The assumption proposed in Gorbunov et al. (2022a), is designed specifically for analyzing vanilla Stochastic EG, it does not have $\{\sigma_k^2\}_{k\geq 0}$ sequence (not able to capture variants of Stochastic EG with variance reduction, quantization, nor coordinate-wise randomization) and works only for (1) with $R(x) \equiv 0$. For more detailed comparison of our approach and this line of work, see Appendix A.

Quasi-strongly monotone case. Under Assumption 2.1 and quasi-strong monotonicity of F, we derive the following general result.

Theorem 2.2. Let F be μ-quasi-strongly monotone ($\mu > 0$) and let Assumption 2.1 hold. Assume that $0 < \gamma \leq \min\left\{1/\mu, 1/2(A+CM)\right\}$ for some $M > B/\rho$ (when $B = 0$, we suppose $M = 0$ and $B/M := 0$ in all following expressions). Then the iterates of SGDA (5), satisfy:

$$\begin{aligned} \mathbb{E}[V_k] &\leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^k V_0 \\ &\quad + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - B/M\right\}}. \qquad (8) \end{aligned}$$

where the Lyapunov function V_k is defined by $V_k = \|x^k - x^{*,k}\|^2 + M\gamma^2\sigma_k^2$ for all $k \geq 0$.

The above theorem states that SGDA (5) converges linearly to the neighborhood of the solution. The size of the neighborhood is proportional to the noises D_1 and D_2. When $D_1 = D_2 = 0$, i.e., the method is variance reduced, it converges linearly to the exact solution in expectation. However, in general, to achieve any predefined accuracy, one needs to reduce the size of the neighborhood somehow. One possible way to do that is to use a proper stepsize schedule. We formalize this discussion in the following result.

Corollary 2.3. Let the assumptions of Theorem 2.2 hold. Consider two possible cases.

Case 1. Let $D_1 = D_2 = 0$. Then, for any $K \geq 0$, $M = 2B/\rho$, and $\gamma = \min\left\{1/\mu, 1/2(A+2BC/\rho)\right\}$, the iterates of SGDA, given by (5), satisfy: $\mathbb{E}[V_K] \leq V_0 \exp\left(-\min\left\{\frac{\mu}{2(A+2BC/\rho)}, \frac{\rho}{2}\right\}K\right)$.

[2] Although Assumption 2.1 does not formally imply star-cocoercivity of F, but in all special cases, considered in this work, operator F is star-cocoercive.

Case 2. Let $D_1 + MD_2 > 0$. For any $K \geq 0$ and $M = {}^{2B}/_{\rho}$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\gamma_k = \frac{1}{h} \quad \text{if } K \leq \frac{h}{\mu} \text{ or } \left(K > \frac{h}{\mu} \text{ and } k < k_0\right),$$
$$\gamma_k = \frac{2}{\mu(\kappa + k - k_0)} \quad \text{if } K > \frac{h}{\mu} \text{ and } k \geq k_0,$$

where $h = \max\{2(A + {}^{2BC}/_{\rho}), {}^{2\mu}/_{\rho}\}$, $\kappa = {}^{2h}/_{\mu}$ and $k_0 = \lceil {}^{K}/_{2} \rceil$. For this choice of γ_k, the iterates of SGDA, given by (5), satisfy:

$$\mathbb{E}[V_K] \leq \frac{32hV_0}{\mu}\exp\left(-\frac{\mu}{h}K\right) + \frac{36(D_1 + {}^{2BD_2}/_{\rho})}{\mu^2 K}.$$

Monotone case. When $\mu = 0$, we additionally assume that F is monotone, i.e., for all $x, y \in \mathbb{R}^d$

$$\langle F(x) - F(y), x - y\rangle \geq 0.$$

Similar to minimization, in the case of $\mu = 0$, the squared distance to the solution is not a valid measure of convergence. To introduce an appropriate convergence measure, we make the following assumption.

Assumption 2.4. There exists a compact convex set $\mathcal{C}$ (with the diameter $\Omega_{\mathcal{C}} := \max_{x,y\in\mathcal{C}} \|x - y\|$) such that $X^* \subset \mathcal{C}$.

In this setting, we focus on the following quantity called a restricted gap-function (Nesterov, 2007) defined for any $z \in \mathbb{R}^d$ and any $\mathcal{C} \subset \mathbb{R}^d$ satisfying Assumption 2.4:

$$\mathrm{Gap}_{\mathcal{C}}(z) := \max_{u\in\mathcal{C}} \left[\langle F(u), z - u\rangle + R(z) - R(u)\right]. \quad (9)$$

Assumption 2.4 and function $\mathrm{Gap}_{\mathcal{C}}(z)$ are standard for the convergence analysis of methods for solving (1) with monotone F (Nesterov, 2007; Alacaoglu and Malitsky, 2021). Additional discussion is left to Appendix D.2.

Under these assumptions, Assumption 2.1, and star-cocoercivity we derive the following general result.

Theorem 2.5. Let F be monotone, ℓ-star-cocoercive and let Assumptions 2.1, 2.4 hold. Assume that $0 < \gamma \leq {}^{1}/_{2(A+{}^{BC}/_{\rho})}$. Then for all $K \geq 0$ the iterates of SGDA, given by (5), satisfy:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \\
&\leq \frac{3\left[\max_{u\in\mathcal{C}} \|x^0 - u\|^2\right]}{2\gamma K} + 9\gamma \max_{x^*\in X^*} \|F(x^*)\|^2 \\
&\quad + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + (4A + \ell + {}^{8BC}/_{\rho}) \cdot \frac{\|x^0 - x^{*,0}\|^2}{K} \\
&\quad + (4 + (4A + \ell + {}^{8BC}/_{\rho})\gamma)\frac{\gamma B\sigma_0^2}{\rho K} \\
&\quad + \gamma(2 + \gamma(4A + \ell + {}^{8BC}/_{\rho}))(D_1 + {}^{2BD_2}/_{\rho}). \quad (10)
\end{aligned}$$

The above result establishes $\mathcal{O}({}^{1}/_{K})$ rate of convergence to the accuracy proportional to the stepsize γ multiplied by the noise term $D_1 + {}^{2BD_2}/_{\rho}$ and $\max_{x^*\in X^*} \|F(x^*)\|^2$. We notice that if $R \equiv 0$ in (1), then $F(x^*) = 0$, meaning that in this case, the second term from (10) equals zero. Otherwise, even in the deterministic case one needs to use small stepsizes to ensure the convergence to any predefined accuracy (see Corollary D.4 in Appendix D.2).

Cocoercive case. The term proportional to $\max_{x^*\in X^*} \|F(x^*)\|^2$ can be removed if we assume that the operator F is not just monotone star-cocoercive (4), but general cocoercive, i.e., it holds that for all $x, y \in \mathbb{R}^d$

$$\|F(x) - F(y)\|^2 \leq \ell\langle F(x) - F(y), x - y\rangle.$$

Theorem 2.6. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4 hold. Assume that $0 < \gamma \leq \min\{{}^{1}/_{\ell}, {}^{1}/_{2(A+{}^{BC}/_{\rho})}\}$. Then for all $K \geq 0$ the iterates of SGDA, given by (5), satisfy:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \\
&\leq \frac{3\max_{u\in\mathcal{C}} \|x^0 - u\|^2}{2\gamma K} \\
&\quad + (6A + 3\ell + {}^{12BC}/_{\rho}) \cdot \frac{\|x^0 - x^{*,0}\|^2}{K} \\
&\quad + (6 + (6A + 3\ell + {}^{12BC}/_{\rho})\gamma)\frac{\gamma B\sigma_0^2}{\rho K} \\
&\quad + \gamma(3 + \gamma(6A + 3\ell + {}^{12BC}/_{\rho}))(D_1 + {}^{2BD_2}/_{\rho}). \quad (11)
\end{aligned}$$

In contrast to Theorem 2.5, the above result implies $\mathcal{O}({}^{1}/_{K})$ convergence rate in the deterministic case. See Corollary D.6 in Appendix D.3 for the results of the convergence with a selected stepsize.

3 SGDA WITH ARBITRARY SAMPLING

We start our consideration of special cases with a standard SGDA (5) with $g^k = F_{\xi^k}(x^k), \xi^k \sim \mathcal{D}$ under so-called *expected cocoercivity* assumption from Loizou et al. (2021), which we properly adjust to the setting of regularized VIPs.

Assumption 3.1 (Expected Cocoercivity)**.** We assume that stochastic operator $F_\xi(x), \xi \sim \mathcal{D}$ is such that for all $x \in \mathbb{R}^d$, $\mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x) - F_\xi(x^*)\|^2\right] \leq \ell_{\mathcal{D}}\langle F(x) - F(x^*), x - x^*\rangle$, where $x^* = \mathrm{proj}_{X^*}(x)$.

When $R(x) \equiv 0$, this assumption recovers the original one from Loizou et al. (2021). We also emphasize that for operator F Assumption 3.1 implies only star-cocoercivity.

Following Loizou et al. (2021), we mainly focus on the finite-sum case and its stochastic reformulation: we consider a random *sampling* vector $\xi = (\xi_1, \ldots, \xi_n)^\top \in \mathbb{R}^n$ having a distribution $\mathcal{D}$ such that $\mathbb{E}_{\mathcal{D}}[\xi_i] = 1$ for all $i \in [n]$. Using this we can rewrite $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$ as

$$F(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{D}}[\xi_i F_i(x)] = \mathbb{E}_{\mathcal{D}}\left[F_\xi(x)\right], \quad (12)$$

where $F_\xi(x) = \frac{1}{n}\sum_{i=1}^n \xi_i F_i(x)$. Such a reformulation allows to handle a wide range of samplings: the only assumption on $\mathcal{D}$ is $\mathbb{E}_{\mathcal{D}}[\xi_i] = 1$ for all $i \in [n]$. Therefore, this setup is often referred to as *arbitrary sampling* (Richtárik and Takác, 2020; Loizou and Richtárik, 2020a,b; Gower et al., 2019, 2021; Hanzely and Richtárik, 2019; Qian et al., 2019, 2021a). We elaborate on several special cases in Appendix E.5.

In this setting, SGDA with Arbitrary Sampling (SGDA-AS)[3] fits our framework.

Proposition 3.2. Let Assumption 3.1 hold. Then, SGDA-AS satisfies Assumption 2.1 with $A = \ell_{\mathcal{D}}$, $D_1 = 2\sigma_*^2 := 2\max_{x^* \in X^*}\mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x^*) - F(x^*)\|^2\right]$, $B = 0$, $\sigma_k^2 \equiv 0$, $C = 0$, $\rho = 1$, $D_2 = 0$.

Plugging these parameters to Theorem 2.2 we recover the result[4] from Loizou et al. (2021) when $R(x) \equiv 0$ and generalize it to the case of $R(x) \not\equiv 0$ without sacrificing the rate. Applying Corollary 2.3, we establish the rate of convergence to the exact solution.

Corollary 3.3. Let F be μ-quasi-strongly monotone and Assumption 3.1 hold. Then for all $K > 0$ there exists a choice of γ (see (48)) for which the iterates of SGDA-AS, satisfy:

$$\mathbb{E}[\|x^K - x^{*,K}\|^2] = \mathcal{O}\left(\frac{\ell_{\mathcal{D}}\Omega_0^2}{\mu}\exp\left(-\frac{\mu}{\ell_{\mathcal{D}}}K\right) + \frac{\sigma_*^2}{\mu^2 K}\right),$$

where $\Omega_0^2 = \|x^0 - x^{*,0}\|^2$.

For the different stepsize schedule, Loizou et al. (2021) derive the convergence rate $\mathcal{O}(1/K + 1/K^2)$ which is inferior to our rate, especially when σ_*^2 is small. In addition, Loizou et al. (2021) consider explicitly only uniform minibatch sampling without replacement as a special case of arbitrary sampling. In Appendix E.5, we discuss another prominent sampling strategy called importance sampling. In Section 6, we provide numerical experiments verifying our theoretical findings and showing the benefits of importance sampling over uniform sampling for SGDA.

4 SGDA WITH VARIANCE REDUCTION

In this section, we focus on variance-reduced variants of SGDA for solving finite-sum problems $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$. We start with the Loopless Stochastic Variance Reduced Gradient Descent-Ascent (L-SVRGDA), which is a generalization of the L-SVRG algorithm proposed in Hofmann et al. (2015); Kovalev et al. (2020). L-SVRGDA (see Alg. 2) follows the update rule (5) with

$$g^k = F_{j_k}(x^k) - F_{j_k}(w^k) + F(w^k),$$
$$w^{k+1} = \begin{cases} x^k, & \text{with prob. } p, \\ w^k, & \text{with prob. } 1-p, \end{cases} \quad (13)$$

where in k^{th} iteration j_k is sampled uniformly at random from $[n]$. Here full operator F is computed once w^k is updated, which happens with probability p. Typically, p is chosen as $p \sim 1/n$ ensuring that the expected cost of 1 iteration equals $\mathcal{O}(1)$ oracle calls, i.e., computations of $F_i(x)$ for some $i \in [n]$.

We introduce the following assumption about operators F_i.

Assumption 4.1 (Averaged Star-Cocoercivity). We assume that there exists a constant $\widehat{\ell} > 0$ such that for all $x \in \mathbb{R}^d$

$$\frac{1}{n}\sum_{i=1}^n \|\Delta_{F_i}(x, x^*)\|^2 \leq \widehat{\ell}\langle F(x) - F(x^*), x - x^*\rangle, \quad (14)$$

where $\Delta_{F_i}(x, x^*) = F_i(x) - F_i(x^*)$ and $x^* = \mathrm{proj}_{X^*}(x)$.

For example, if F_i is ℓ_i-cocoercive for $i \in [n]$, then (14) holds with $\widehat{\ell} \leq \max_{i\in[n]} \ell_i$. Next, if F_i is L_i-Lipschitz for all $i \in [n]$ and F is μ-quasi strongly monotone, then (14) is satisfied for $\widehat{\ell} \in [\overline{L}, \overline{L}^2/\mu]$, where $\overline{L}^2 = \frac{1}{n}\sum_{i=1}^n L_i^2$.

Moreover, for the analysis of variance-reduced variants of SGDA we also use the uniqueness of the solution.

Assumption 4.2 (Unique Solution). We assume that the solution set X^* of problem (1) is a singleton: $X^* = \{x^*\}$.

These assumptions are sufficient to derive validity of Assumption 2.1 for L-SVRGDA estimator.

Proposition 4.3. Let Assumptions 4.1 and 4.2 hold. Then, L-SVRGDA satisfies Assumption 2.1 with $A = \widehat{\ell}$, $B = 2$, $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^*)\|^2$, $C = p\widehat{\ell}/2$, $\rho = p$, $D_1 = D_2 = 0$.

Plugging these parameters in our general results on the convergence of SGDA-type algorithms we derive the conver-

[3] For the pseudo-code of SGDA-AS see Algorithm 1 in Appendix E.

[4] In the main part of the paper, we focus on μ-quasi strongly monotone case with $\mu > 0$. For simplicity, we provide here the rates of convergence to the exact solution. Further details, including the rates in monotone case, are left to the Appendix.

gence results for L-SVRGDA, see Table 1 and Appendix F.1 for the details. Moreover, in Appendix F.2, we show that SAGA-SGDA (Palaniappan and Bach, 2016) fits our framework and using our general analysis we tighten the convergence rates for this method.

We compare our convergence guarantees with known results in Table 1. We note that by neglecting importance sampling scenario, in the worst case, our convergence results match the best-known results for SGDA-type methods, i.e., ones derived in Palaniappan and Bach (2016). Indeed, this follows from $\widehat{\ell} \in [\overline{L}, \overline{L}^2/\mu]$. Next, when the difference between $\overline{\ell}$ and $\widehat{\ell}$ is not significant, our complexity results match the one derived in Chavdarova et al. (2019) for SVRE, which is EG-type method. Although in general, $\overline{\ell}$ might be smaller than $\widehat{\ell}$, our analysis does not require cocoercivity of each F_i and it works for $R(x) \not\equiv 0$. Finally, Alacaoglu and Malitsky (2021) derive a better rate (when $n = \mathcal{O}(\overline{L}^2/\mu^2)$), but their method is based on EG. Therefore, our results match the best-known ones in the literature on SGDA-type methods.

5 DISTRIBUTED SGDA WITH COMPRESSION

In this section, we consider the distributed version of (1), i.e., we assume that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$, where $\{F_i\}_{i=1}^n$ are distributed across n devices connected with parameter-server in a centralized fashion. Each device i has an access to the computation of the unbiased estimate of F_i at the given point. Typically, in these settings, communication is a bottleneck, especially when n and d are huge. This means that in the naive distributed implementations of SGDA, communication rounds take much more time than local computations on the clients. Various approaches are used to circumvent this issue.

One of them is based on the usage of compressed communications. We focus on unbiased compression operators.

Definition 5.1. Operator $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ (possibly randomized) is called *unbiased compressor/quantization* if there exists a constant $\omega \geq 0$ such that for all $x \in \mathbb{R}^d$

$$\mathbb{E}[\mathcal{Q}(x)] = x, \quad \mathbb{E}[\|\mathcal{Q}(x) - x\|^2] \leq \omega\|x\|^2. \tag{15}$$

In this paper, we consider compressed communications in the direction from clients to the server. The simplest method with compression – QSGDA (Alg. 4) – can be described as SGDA (5) with $g^k = \frac{1}{n}\sum_{i=1}^n \mathcal{Q}(g_i^k)$. Here g_i^k are stochastic estimators satisfying the following assumption[5].

[5]We use this assumption for illustrating the flexibility of the framework. It is possible to consider Arbitrary Sampling setup as well.

Assumption 5.2 (Bounded variance)**.** All stochastic realizations g_i^k are unbiased and have bounded variance, i.e., for all $i \in [n]$ and $k \geq 0$ the following holds:

$$\mathbb{E}[g_i^k] = F_i(x^k), \quad \mathbb{E}[\|g_i^k - F_i(x^k)\|^2] \leq \sigma_i^2. \tag{16}$$

Despite its simplicity, QSGDA was never considered in the literature on solving min-max problems and VIPs. It turns out that under such assumptions QSGDA satisfies our Assumption 2.1.

Proposition 5.3. Let F be ℓ-star-cocoercive and Assumptions 4.1, 5.2 hold. Then, QSGDA satisfies Assumption 2.1 with $A = \frac{3\ell}{2} + \frac{9\omega\widehat{\ell}}{2n}$, $B = 0$, $\sigma_k^2 \equiv 0$, $D_1 = \frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n}$, $C = 0$, $\rho = 1$, $D_2 = 0$, where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2, \zeta_*^2 := \frac{1}{n}\max_{x*\in X^*}\sum_{i=1}^n \|F_i(x^*)\|^2$.

As for the other special cases, we derive the convergence results for QSGDA using our general theorems (see Table 2 and Appendix G.1 for the details). The proposed method is simple, but has a significant drawback: even in the deterministic case ($\sigma = 0$), QSGDA does not converge linearly unless $\zeta_*^2 = 0$. However, when the data on clients is arbitrarily heterogeneous the dissimilarity measure ζ_*^2 is strictly positive and can be large (even when $R(x) \equiv 0$).

To resolve this issue, we propose a more advanced scheme based on DIANA update (Mishchenko et al., 2019; Horváth et al., 2019) – DIANA-SGDA (Alg. 5). In a nutshell, DIANA-SGDA is SGDA (5) with g^k defined as follows:

$$\begin{aligned}
\Delta_i^k &= g_i^k - h_i^k, \quad h_i^{k+1} = h_i^k + \alpha\mathcal{Q}(\Delta_i^k), \\
g^k &= h^k + \frac{1}{n}\sum_{i=1}^n \mathcal{Q}(\Delta_i^k), \\
h^{k+1} &= \frac{1}{n}\sum_{i=1}^n h_i^{k+1} = h^k + \alpha\frac{1}{n}\sum_{i=1}^n \mathcal{Q}(\Delta_i^k),
\end{aligned} \tag{17}$$

where the first two lines correspond to the local computations on the clients and the last two lines – to the server-side computations. Taking into account the update rule for h^{k+1}, one can notice that DIANA-SGDA requires workers to send only vectors $\mathcal{Q}(\Delta_i^k)$ to the server at step k, i.e., the method uses only compressed workers-server communications.

As we show next, DIANA-SGDA fits our framework.

Proposition 5.4. Let Assumptions 4.1, 4.2, 5.2 hold. Suppose that $\alpha \leq 1/(1+\omega)$. Then, DIANA-SGDA with quantization (15) satisfies Assumption 2.1 with $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - F_i(x^*)\|^2$ and $A = \left(\frac{1}{2} + \frac{\omega}{n}\right)\widehat{\ell}$, $B = \frac{2\omega}{n}$, $D_1 = \frac{(1+\omega)\sigma^2}{n}$, $C = \frac{\alpha\widehat{\ell}}{2}$, $\rho = \alpha$, $D_2 = \alpha\sigma^2$, where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

DIANA-SGDA can be considered as a variance-reduced method since it reduces the term proportional to $\omega\zeta_*^2$ that the bound for QSGDA contains (see Table 2 and Appendix G.2

Table 1: Summary of the complexity results for variance reduced methods for solving (1). By complexity we mean the number of oracle calls required for the method to find x such that $\mathbb{E}[\|x - x^*\|^2] \le \varepsilon$. Dependencies on numerical and logarithmic factors are hidden. By default, operator F is assumed to be μ**-strongly monotone** and, as the result, the solution is unique. Our results rely on μ**-quasi strong monotonicity** of F (3), but we also assume uniqueness of the solution. Methods supporting $R(x) \not\equiv 0$ are highlighted with *. Our results are highlighted in green. Notation: $\overline{\ell}$, $\overline{L}$ = averaged cocoercivity/Lipschitz constants depending on the sampling strategy, e.g., for uniform sampling $\overline{\ell}^2 = \frac{1}{n}\sum_{i=1}^n \ell_i^2$, $\overline{L}^2 = \frac{1}{n}\sum_{i=1}^n L_i^2$ and for importance sampling $\overline{\ell} = \frac{1}{n}\sum_{i=1}^n \ell_i$, $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$; $\widehat{\ell}$ = averaged star-cocoercivity constant from Assumption 4.1.

Method	Citation	Assumptions	Complexity
SVRE [(1)]	(Chavdarova et al., 2019)	F_i is ℓ_i-cocoer.	$n + \frac{\overline{\ell}}{\mu}$
EG-VR *[(1)]	(Alacaoglu and Malitsky, 2021)	F_i is L_i-Lip.	$n + \sqrt{n}\frac{\overline{L}}{\mu}$
SVRGDA *	(Palaniappan and Bach, 2016)	F_i is L_i-Lip.	$n + \frac{\overline{L}^2}{\mu^2}$
SAGA-SGDA *	(Palaniappan and Bach, 2016)	F_i is L_i-Lip.	$n + \frac{\overline{L}^2}{\mu^2}$
VR-AGDA	(Yang et al., 2020)	F_i is $L_{\max}$-Lip.[(2)]	$\min\left\{n + \frac{L_{\max}^9}{\mu^9}, n^{2/3}\frac{L_{\max}^3}{\mu^3}\right\}$
L-SVRGDA *	**This paper**	As. 4.1	$n + \frac{\widehat{\ell}}{\mu}$
SAGA-SGDA *	**This paper**	As. 4.1	$n + \frac{\widehat{\ell}}{\mu}$

[(1)] The method is based on Extragradient update rule.
[(2)] Yang et al. (2020) consider saddle point problems satisfying so-called two-sided PL condition, which is weaker than strong-convexity-strong-concavity of the objective function.

for the details). As the result, when $\sigma = 0$, i.e., workers compute $F_i(x)$ at each step, DIANA-SGDA enjoys linear convergence to the exact solution.

Next, when local operators F_i have a finite-sum form $F_i(x) = \frac{1}{m}\sum_{j=1}^m F_{ij}(x)$, one can combine L-SVRGDA and DIANA-SGDA as follows: consider the scheme from (17) with

$$g_i^k = F_{ij_k}(x^k) - F_{ij_k}(w^k) + F(w_i^k),$$
$$w_i^{k+1} = \begin{cases} x^k, & \text{with prob. } p, \\ w_i^k, & \text{with prob. } 1-p, \end{cases} \tag{18}$$

where j_k is sampled uniformly at random from $[m]$. We call the resulting method VR-DIANA-SGDA (Alg. 6) and we note that its analog for solving minimization problems (VR-DIANA) was proposed and analyzed in Horváth et al. (2019).

To cast VR-DIANA-SGDA as a special case of our general framework, we need to make the following assumption.

Assumption 5.5. We assume that there exists a constant $\widetilde{\ell} > 0$ such that for all $x \in \mathbb{R}^d$

$$\frac{1}{nm}\sum_{i,j=1,1}^{n,m} \|\Delta_{F_{ij}}(x,x^*)\|^2 \le \widetilde{\ell}\langle F(x) - F(x^*), x - x^*\rangle, \tag{19}$$

where $\Delta_{F_{ij}}(x,x^*) = F_{ij}(x) - F_{ij}(x^*)$, $x^* = \mathrm{proj}_{X^*}(x)$.

Using Assumption 5.5 and previously introduced conditions, we get the following result.

Proposition 5.6. Let F be ℓ-star-cocoercive and Assumptions 4.1, 4.2, 5.5 hold. Suppose that $\alpha \le \min\left\{\frac{p}{3}, \frac{1}{1+\omega}\right\}$. Then, VR-DIANA-SGDA satisfies Assumption 2.1 with $A = \frac{\ell}{2} + \frac{\widetilde{\ell}}{n} + \frac{\omega(\widehat{\ell}+\widetilde{\ell})}{n}$, $B = \frac{2(\omega+1)}{n}$, $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - F_i(x^*)\|^2 + \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|F_{ij}(w_i^k) - F_{ij}(x^*)\|^2$, $C = \frac{p\widetilde{\ell}}{2} + \alpha(\widetilde{\ell} + \widehat{\ell})$, $\rho = \alpha$, $D_1 = D_2 = 0$.

Since $D_1 = D_2 = 0$, our general results imply linear convergence of VR-DIANA-SGDA when $\mu > 0$ (see the details in Appendix G.3). That is, VR-DIANA-SGDA is the *first linearly converging distributed* SGDA*-type method with compression*. We compare it with MASHA1 (Beznosikov et al., 2021b) in Table 2. Firstly, let us note that MASHA1 is a method based on EG, and its convergence guarantees depend on the Lipschitz constants. In addition, we note that the complexity of MASHA1 could be better than the one of VR-DIANA-SGDA when cocoercivity constants are large compared to Lipschitz ones. However, our compleixty bound has better dependency on quantization parameter ω, number of clients n, and the size of the local dataset m. These parameters can be large meaning that the improvement is noticeable.

6 NUMERICAL EXPERIMENTS

To illustrate our theoretical results, we conduct several numerical experiments on quadratic games, which are defined through the affine operator: $F(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i x + b_i$, where each matrix $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ is non-symmetric with all eigenvalues having strictly positive real parts. Enforcing all the eigenvalues to have strictly positive real part ensures that the operator is strongly monotone and cocoercive. We consider two different settings: (i) problem without constraints, and (ii) problem that has ℓ_1 regularization and constraints forcing the solution to lie in the ℓ_∞-ball of radius r. In all experiments, we use a constant stepsize for all methods

Table 2: Summary of the complexity results for distributed methods with unbiased compression for solving distributed (1) with $F = \frac{1}{n}\sum_{i=1}^n F_i(x)$. By complexity we mean the number of communication rounds required for the method to find x such that $\mathbb{E}[\|x - x^*\|^2] \le \varepsilon$. Dependencies on numerical and logarithmic factors are hidden. $\mathbb{E}$ stands for the setup, when $F_i(x) = \mathbb{E}_{\xi_i}[F_{\xi_i}(x)]$; Σ denotes the case, when $F_i(x) = \frac{1}{m}\sum_{j=1}^m F_{ij}(x)$. Our results rely on μ**-quasi strong monotonicity** of F (3), but we also assume the uniqueness of the solution. Methods supporting $R(x) \not\equiv 0$ are highlighted with *. Our results are highlighted in green. Notation: $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ – averaged upper bound for the variance (see Ass. 5.2 for the definition of σ_i^2); ω = quantization parameter (see Def. 5.1); $\zeta_*^2 = \frac{1}{n}\max_{x^*\in X^*}\sum_{i=1}^n \|F_i(x^*)\|^2$; $L_{\max} = \max_{i\in[n]} L_i$; $\widetilde{\ell}$ = averaged star-cocoercivity constant from Ass. 5.5.

Setup	Method	Citation	Assumptions	Complexity
$\mathbb{E}$	QSGDA *	**This paper**	As. 4.1, 5.2	$\frac{\ell}{\mu} + \frac{\omega\widehat{\ell}}{n\mu} + \frac{(1+\omega)\sigma^2 + \omega\zeta_*^2}{n\mu^2\varepsilon}$
	DIANA-SGDA *	**This paper**	As. 4.1, 5.2	$\omega + \frac{\ell}{\mu} + \frac{\omega\widehat{\ell}}{n\mu} + \frac{(1+\omega)\sigma^2}{n\mu^2\varepsilon}$
Σ	MASHA1 $^{*(1)}$	(Beznosikov et al., 2021b)	F_i is L_i-Avg. Lip.[(2)]	$m + \omega + \frac{L_{\max}\sqrt{(m+\omega)\left(1+\frac{\omega}{n}\right)}}{\mu}$
	VR-DIANA-SGDA *	**This paper**	As. 4.1, 5.5	$m + \omega + \frac{\ell}{\mu} + \frac{(1+\omega)(\widehat{\ell}+\widetilde{\ell})}{n\mu} + \frac{(1+\omega)\max\{m,\omega\}\widetilde{\ell}}{nm\mu}$

[(1)] The method is based on Extragradient update rule.
[(2)] This means that for all $x, y \in \mathbb{R}^d$ and $i \in [n]$ the following inequality holds: $\frac{1}{m}\sum_{j=1}^m \|F_{ij}(x) - F_{ij}(y)\|^2 \le L_i^2\|x-y\|^2$.

which was selected manually using a grid search and picking the best-performing stepsize for each method. For further details about the experiments and additional experiments see Appendix B.

Uniform sampling (US) **vs Important sampling** (IS)**.** We note that Loizou et al. (2021) which studies SGDA-AS does not consider IS explicitly. Although we show the theoretical benefits of IS in comparison to US in Appendix E.5, here we provide a numerical comparison to illustrate the superiority of IS (on both constrained and unconstrained quadratic games). We choose the matrices $\mathbf{A}_i$ such that $\ell_{\max} = \max_i \ell_i \gg \bar{\ell}$. In this case, our theory predicts that IS should perform better than US. We provide the results in Fig. 1. We observe that indeed SGDA with IS converges faster and to a smaller neighborhood than SGDA with US. This observation perfectly corroborates our theory.

Comparison of variance reduced methods. In this experiment, we test the performance of our proposed L-SVRGDA (Alg. 2) and compare it to other variance-reduced methods on quadratic games, see Fig. 2. In particular, we compare it to SVRG (Palaniappan and Bach, 2016), SVRE (Chavdarova et al., 2019), EG-VR (Alacaoglu and Malitsky, 2021) and VR-AGDA (Yang et al., 2020). In the constrained setting, we only compare L-SVRGDA to SVRG and EG-VR, since they are the only methods from this list that handle constrained settings. For loopless variants, we choose $p = \frac{1}{n}$ and for the non-loopless variants we pick the number of inner-loop iterations to be n. We observe that all methods converge linearly and that L-SVRGDA is competitive with the other considered variance-reduced methods, converging slightly faster than all of them.

We point out that we plot the distance to optimality as a function of the number of oracle calls. When using variance-reduced methods we sometimes have to compute the full-batch gradient, and thus have to make n oracle calls. This is why we observe "steps" for variance-reduced methods in Fig. 2: we observe a "step" every time the full batch gradient is computed.

Comparison of distributed methods. In our last experiment, we consider a distributed version of the quadratic game, in which we assume that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$ with each $\{F_i\}_{i=1}^n$ being constructed similarly to the previous experiments. The information about operator F_i is stored on node i only. We compare the distributed methods proposed in the paper: QSGDA, DIANA-SGDA, and VR-DIANA-SGDA. For the quantization, we use the RandK sparsification (Beznosikov et al., 2020a) with $K = 5$. We show our findings in Fig. 3, where the performance is measured both in terms of the number of oracle calls and the number of bits communicated from workers to the server. In both figures, we can clearly see the advantage of using quantization in terms of reducing the communication cost compared to the baseline SGDA. We also observe that VR-DIANA-SGDA achieves linear convergence to the solution. Additional experiments are deferred to Appendix B.

Acknowledgments

This research of A. Beznosikov has been supported by The Analytical Center for the Government of the Russian Federation (Agreement No. 70-2021-00143 dd. 01.11.2021, IGK 000000D730321P5Q0002).

References

Alacaoglu, A. and Malitsky, Y. (2021). Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*.

Alacaoglu, A., Malitsky, Y., and Cevher, V. (2021). Forward-reflected-backward method with variance reduction. *Computational optimization and applications*, 80(2):321–346.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic,

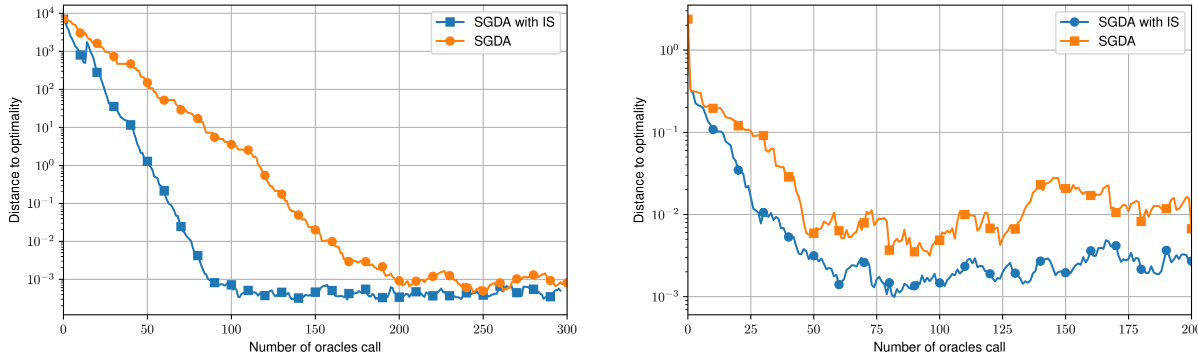

Figure 1: Comparison of Uniform Sampling (US) vs Importance Sampling (IS). **Left:** the result for the problem without constraints, **right:** with constraints. As expected by theory IS converges faster and to a smaller neighborhood than US.

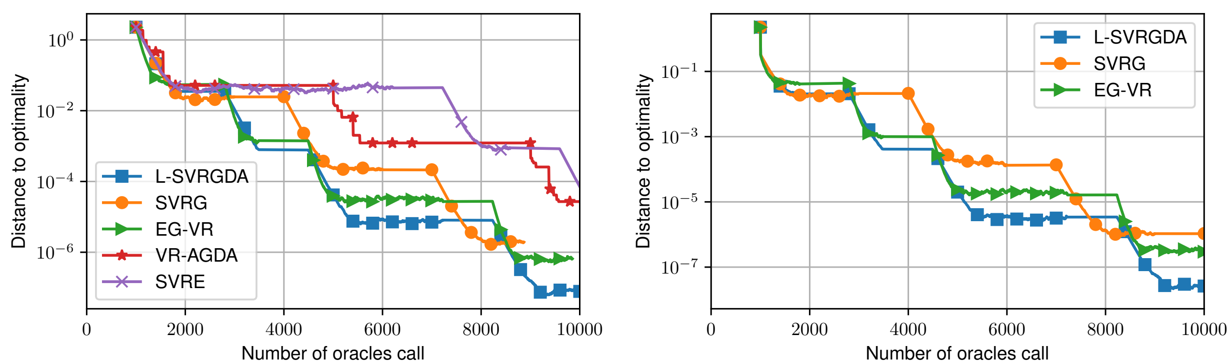

Figure 2: Comparison of variance-reduced methods. **Left:** the result for the problem without constraints, **right:** with constraints. Note that L-SVRGDA is very competitive, and outperforms all the other methods.

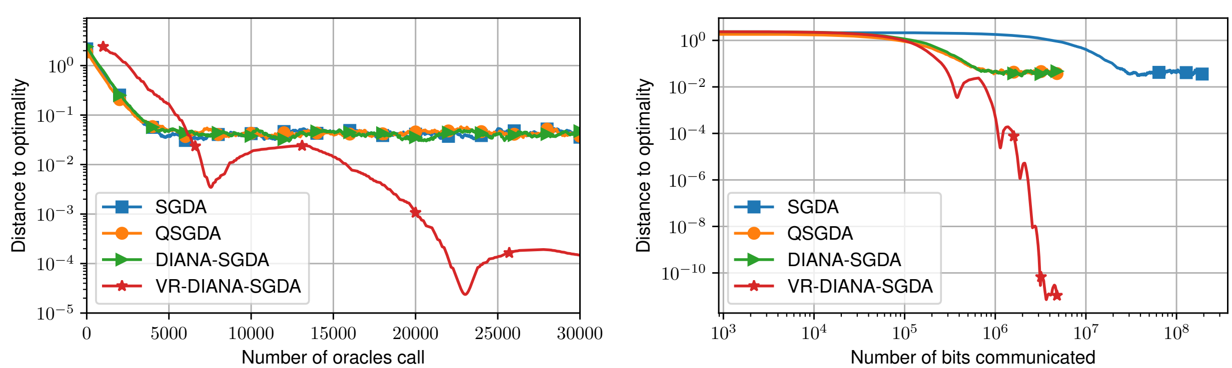

Figure 3: Comparison of algorithms in distributed setting. **Left:** number of oracle calls, **right:** number of bits communicated.

M. (2017). Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720.

Azizian, W., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2021). The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *Conference on Learning Theory*, pages 326–358. PMLR.

Bach, F. (2019). The "η-trick" or the effectiveness of reweighted least-squares.

Bauschke, H. H., Combettes, P. L., et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer.

Beck, A. (2017). *First-order methods in optimization*. Society for Industrial and Applied Mathematics (SIAM).

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2020a). On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*.

Beznosikov, A., Novitskii, V., and Gasnikov, A. (2021a).

One-point gradient-free methods for smooth and non-smooth saddle-point problems. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 144–158. Springer.

Beznosikov, A., Richtárik, P., Diskin, M., Ryabinin, M., and Gasnikov, A. (2021b). Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. *arXiv preprint arXiv:2110.03313*.

Beznosikov, A., Sadiev, A., and Gasnikov, A. (2020b). Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 105–119. Springer.

Beznosikov, A., Samokhin, V., and Gasnikov, A. (2020c). Distributed saddle-point problems: Lower bounds, optimal algorithms and federated gans. *arXiv preprint arXiv:2010.13112*.

Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5):877–905.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. (2019). Variance reduction for matrix games. *Advances in Neural Information Processing Systems*, 32.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. (2019). Reducing noise in GAN training with variance reduced extragradient. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Daskalakis, C., Skoulakis, S., and Zampetakis, M. (2021). The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478.

Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Dem'yanov, V. F. and Pevnyi, A. B. (1972). Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52.

Diakonikolas, J., Daskalakis, C., and Jordan, M. (2021). Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2019). A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gorbunov, E., Berard, H., Gidel, G., and Loizou, N. (2022a). Stochastic Extragradient: General Analysis and Improved Rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR.

Gorbunov, E., Burlachenko, K. P., Li, Z., and Richtarik, P. (2021). MARINA: Faster non-convex distributed learning with compression. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3788–3798. PMLR.

Gorbunov, E., Hanzely, F., and Richtarik, P. (2020a). A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 680–690. PMLR.

Gorbunov, E., Kovalev, D., Makarenko, D., and Richtarik, P. (2020b). Linearly converging error compensated sgd. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc.

Gorbunov, E., Loizou, N., and Gidel, G. (2022b). Extragradient Method: $\mathrm{O}(1/k)$ Last-Iterate Convergence for Monotone variational Inequalities and Connections with Cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR.

Gower, R., Sebbouh, O., and Loizou, N. (2021). Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General Analysis and Improved Rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209.

Han, Y., Xie, G., and Zhang, Z. (2021). Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*.

Hanzely, F., Mishchenko, K., and Richtárik, P. (2018). SEGA: Variance reduction via gradient sketching. *Advances in Neural Information Processing Systems*, 31.

Hanzely, F. and Richtárik, P. (2019). Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 304–312. PMLR.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. (2015). Variance reduced stochastic gradient descent with neighbors. *Advances in Neural Information Processing Systems*, 28.

Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., and Richtárik, P. (2019). Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2019). On the convergence of single-call stochastic extra-gradient methods. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2020). Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26.

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

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. (2019). Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR.

Khaled, A., Sebbouh, O., Loizou, N., Gower, R. M., and Richtárik, P. (2020). Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *arXiv preprint arXiv:2006.11573*.

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

Kovalev, D., Horváth, S., and Richtárik, P. (2020). Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*.

Li, C. J., Yu, Y., Loizou, N., Gidel, G., Ma, Y., Roux, N. L., and Jordan, M. I. (2021). On the convergence of stochastic extragradient for bilinear games with restarted iteration averaging. *arXiv preprint arXiv:2107.00464*.

Li, Z., Kovalev, D., Qian, X., and Richtarik, P. (2020). Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR.

Lin, H., Mairal, J., and Harchaoui, Z. (2018). Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(1):7854–7907.

Lin, T., Zhou, Z., Mertikopoulos, P., and Jordan, M. (2020). Finite-time last-iterate convergence for multi-agent learning in games. In *International Conference on Machine Learning*, pages 6161–6171. PMLR.

Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O'Reilly, U.-M. (2020). Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International Conference on Machine Learning*, pages 6282–6293. PMLR.

Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. (2021). Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34.

Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. (2020). Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR.

Loizou, N. and Richtárik, P. (2020a). Convergence analysis of inexact randomized iterative methods. *SIAM Journal on Scientific Computing*, 42(6):A3979–A4016.

Loizou, N. and Richtárik, P. (2020b). Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710.

Luo, L., Xie, G., Zhang, T., and Zhang, Z. (2021). Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*.

Malitsky, Y. and Tam, M. K. (2020). A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472.

Mertikopoulos, P. and Zhou, Z. (2019). Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. (2019). Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtarik, P., and Malitsky, Y. (2020). Revisiting stochastic extragradient. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial*

Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, pages 4573–4582. PMLR.

Morgenstern, O. and Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

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

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259.

Palaniappan, B. and Bach, F. (2016). Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424.

Popov, L. D. (1980). 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.

Qian, X., Qu, Z., and Richtárik, P. (2019). Saga with arbitrary sampling. In *International Conference on Machine Learning*, pages 5190–5199. PMLR.

Qian, X., Qu, Z., and Richtárik, P. (2021a). L-svrg and l-katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22(112):1–47.

Qian, X., Richtárik, P., and Zhang, T. (2021b). Error compensated distributed sgd can be accelerated. *Advances in Neural Information Processing Systems*, 34.

Richtárik, P., Sokolov, I., and Fatkhullin, I. (2021). EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems*.

Richtárik, P. and Takác, M. (2020). Stochastic reformulations of linear systems: algorithms and convergence theory. *SIAM Journal on Matrix Analysis and Applications*, 41(2):487–524.

Sadiev, A., Beznosikov, A., Dvurechensky, P., and Gasnikov, A. (2021). Zeroth-order algorithms for smooth saddle-point problems. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 71–85. Springer.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. (2014). 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. (2020). Optimistic dual extrapolation for coherent non-monotone variational inequalities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14303–14314. Curran Associates, Inc.

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

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified sgd with memory. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4452–4463.

Tominin, V., Tominin, Y., Borodich, E., Kovalev, D., Gasnikov, A., and Dvurechensky, P. (2021). On accelerated methods for saddle-point problems with composite structure. *arXiv preprint arXiv:2103.09344*.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR.

Vũ, B. C. (2013). A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. (2020). Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. (2017). Terngrad: ternary gradients to reduce communication in distributed deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1508–1518.

Yang, J., Kiyavash, N., and He, N. (2020). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1153–1165. Curran Associates, Inc.

Yoon, T. and Ryu, E. K. (2021). Accelerated algorithms for smooth convex-concave minimax problems with $\mathrm{O}(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pages 12098–12109. PMLR.

Yuan, D., Ma, Q., and Wang, Z. (2014). Dual averaging method for solving multi-agent saddle-point problems with quantized information. *Transactions of the Institute of Measurement and Control*, 36(1):38–46.

Zhu, D. L. and Marcotte, P. (1996). Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726.

Stochastic Gradient Descent-Ascent: Unified Theory and New Efficient Methods Supplementary Materials

Contents

A FURTHER RELATED WORK

The references necessary to motivate our work and connect it to the most relevant literature are included in the appropriate sections of the main body of the paper. Here we present a broader view of the literature, including some more references to papers of the area that are not directly related with our work.

Variants of the key assumption in prior work & Detailed comparison to our results. Here we would like to provide more details on the comparison with the closely related works (Gorbunov et al., 2020a, 2022a; Loizou et al., 2021).

As we mention in the main part of the paper, Gorbunov et al. (2020a) focus on solving the much simpler minimization problems using SGD. In particular, their Assumption 4.1 requires a function suboptimality (or Bregman divergence) for the upper bound, a concept that cannot be used in VI problems (there are no functions). Thus, the difference of the two notions does not solely lie on the norm bound, but begins at the deeper, conceptual level. In addition, we focus also on monotone VIs (non-quasi-strongly monotone), while Gorbunov et al. (2020a) consider only the class of quasi-strongly convex minimization problems.

Next, Gorbunov et al. (2022a) provide convergence guarantees for vanilla SEG under the arbitrary sampling paradigm. Their analysis is not able to capture SEG with variance reduction, quantization, and coordinate-wise randomization. In contrast, our approach covers variants of SGDA with variance reduction, quantization and coordinate-wise randomization. We are able to capture these more advanced variants by using sequence $\{\sigma_k^2\}_{k\geq 0}$ (see (7)) in our key assumption, and this is a major difference between our approach and the approach of Gorbunov et al. (2022a). In addition, our analysis works for the case $R(x) \not\equiv 0$. Although the generalization of the analysis to the case of non-zero R might be trivial in the quasi-strongly monotone case, for the monotone case this is definitely not straightforward. Finally, for the monotone case, we do not require large batch-sizes to achieve any predefined accuracy, while analysis of SEG in (Gorbunov et al., 2022a) does (see Appendix B in their work).

Finally, we highlight again that Loizou et al. (2021) focus only on uniform minibatch SGDA for solving quasi-strongly monotone problems. This is only a special case of our approach (see Section 3). We note that even in this scenario, through our analysis we were able to provide faster convergence by considering SGDA with importance sampling (see Appendix E.5 and Fig. 1).

Stochastic methods for solving VIPs. Although this paper is devoted to SGDA-type methods, we briefly mention here the works studying other popular stochastic methods for solving VIPs based on different algorithmic schemes such as Extragradient (EG) method (Korpelevich, 1976) and Optimistic Gradient (OG) method (Popov, 1980). The first analysis of Stochastic EG for solving (quasi-strongly) monotone VIPs was proposed in Juditsky et al. (2011) and then was extended and generalized in various ways (Mishchenko et al., 2020; Hsieh et al., 2020; Beznosikov et al., 2020c; Li et al., 2021; Gorbunov et al., 2022a). Stochastic OG was studied in Gidel et al. (2019); Hsieh et al. (2019); Azizian et al. (2021). In addition, lightweight second-order methods like stochastic Hamiltonian methods and stochastic consensus optimization were studied in Loizou et al. (2020), and Loizou et al. (2021), respectively.

Analysis of SGDA. SGDA is usually analyzed under uniformly bounded variance assumption. That is, $\mathbb{E}[\|g^k - F(x^k)\|^2 \mid x^k] \leq \sigma^2$ is typically assumed to get convergence guarantees (Nemirovski et al., 2009; Mertikopoulos and Zhou, 2019; Yang et al., 2020). This assumption rarely holds, especially for unconstrained VIPs: it is easy to construct an example of (1) with F being a finite sum of linear operators such that the variance is unbounded. Lin et al. (2020) provide a convergence analysis of SGDA under a relative random noise assumption allowing to handle some special cases not covered by uniformly bounded variance assumption. However, relative noise is also a quite strong assumption and usually requires a special type of noise appearing in coordinate methods[6] or in the training of overparameterized models (Vaswani et al., 2019). In their recent work, Loizou et al. (2021) proposed a new weak condition called expected cocoercivity. This assumption fits our theoretical framework (see Section 3) and does not imply strong conditions on the variance of the stochastic estimator but it is stronger than star-cocoercivity of operator F.

Variance reduction for VIPs. The first variance-reduced variants of SGDA (SVRGDA and SAGA-SGDA – analogs of SVRG (Johnson and Zhang, 2013) and SAGA (Defazio et al., 2014)) for solving (1) with strongly monotone operator F having a finite-sum form with Lipschitz summands were proposed in Palaniappan and Bach (2016). For two-sided

[6] For example, see inequality (64) from Appendix H in the case when there is no regularization term, i.e., when $R(x) \equiv 0$ and, as a result, $F(x^*) = 0$ for all $x^* \in X^*$.

PL min-max problems without regularization Yang et al. (2020) proposed a variance-reduced version of SGDA with alternating updates. Since the considered class of problems includes non-strongly-convex-non-strongly-concave min-max problems, the rates from Yang et al. (2020) are inferior to Palaniappan and Bach (2016). There are also several works studying variance-reduced methods based on different methods rather than SGDA. Chavdarova et al. (2019) proposed a combination of SVRG and Extragradient (EG) (Korpelevich, 1976) called SVRE and analyzed the method for strongly monotone VIPs without regularization and with cocoercive summands F_i. The cocoercivity assumption was relaxed to averaged Lipschitzness in Alacaoglu and Malitsky (2021), where the authors proposed another variance-reduced version of EG (EG-VR) based on Loopless variant of SVRG (Hofmann et al., 2015; Kovalev et al., 2020). Loizou et al. (2020) studied stochastic Hamiltonian gradient descent (SHGD), and propose the first stochastic variance reduced Hamiltonian method, named L-SVRHG, for solving stochastic bilinear games and and stochastic games satisfying a "sufficiently bilinear" condition. Moreover, Loizou et al. (2020) provided the first set of global non-asymptotic last-iterate convergence guarantees for a stochastic game over a non-compact domain, in the absence of strong monotonicity assumptions.

We should highlight that the rates from Alacaoglu and Malitsky (2021) match the lower bounds from Han et al. (2021). Under additional assumptions similar results were achieved in Carmon et al. (2019). Alacaoglu et al. (2021) developed variance-reduced method (FoRB-VR) based on Forward-Reflected-Backward algorithm (Malitsky and Tam, 2020), but the derived rates are inferior to those from Alacaoglu and Malitsky (2021).

Using Catalyst acceleration framework of Lin et al. (2018), Palaniappan and Bach (2016); Tominin et al. (2021) achieve (neglecting extra logarithmic factors) similar rates as in Alacaoglu and Malitsky (2021) and Luo et al. (2021) derive even tighter rates for min-max problems. However, as all Catalyst-based approaches, these methods require solving an auxiliary problem at each iteration, which reduces their practical efficiency.

Communication compression for VIPs. While distributed methods with compression were extensively studied for solving minimization problems both for unbiased compression operators (Alistarh et al., 2017; Wen et al., 2017; Mishchenko et al., 2019; Horváth et al., 2019; Li et al., 2020; Khaled et al., 2020; Gorbunov et al., 2021) and biased compression operators (Seide et al., 2014; Stich et al., 2018; Karimireddy et al., 2019; Beznosikov et al., 2020a; Gorbunov et al., 2020b; Qian et al., 2021b; Richtárik et al., 2021), much less is known for min-max problems and VIPs. To the best of our knowledge, the first work on distributed methods with compression for min-max problems is Yuan et al. (2014), where the authors proposed a distributed version of Dual Averaging (Nesterov, 2009) with rounding and showed a convergence to the neighborhood of the solution that cannot be reduced via standard tricks like increasing the batchsize or decreasing the stepsize. More recently, Beznosikov et al. (2021b) proposed new distributed variants of EG with unbiased/biased compression for solving (1) with (strongly) monotone and Lipschitz operator F. Beznosikov et al. (2021b) obtained the first linear convergence guarantees on distributed VIPs with compressed communication.

On quasi-strong monotonicity and star-cocoercivity. In this work we focus on quasi-strongly monotone VI problems, a class of structured non-monotone operators for which we are able to provide tight convergence guarantees and avoid the standard issues (cycling and divergence of the methods) appearing in the more general non-monotone regime.

Since in general non-monotone problems, finding approximate first-order locally optimal solutions is intractable (Daskalakis et al., 2021; Diakonikolas et al., 2021), it is reasonable to consider class of problems that satisfy special structural assumptions on the objective function for which these intractability barriers can be bypassed. Examples of problems belong in this category are the ones of our work which satisfy (3) or, for example, the two-sided PL condition (Yang et al., 2020) or the error-bound condition (Hsieh et al., 2020). It is worth highlighting that quasi-strong monotone problems were considered in Mertikopoulos and Zhou (2019); Song et al. (2020); Loizou et al. (2021); Gorbunov et al. (2022a) as well.

Cocoercivity is a classical assumption in the literature on VIPs (Zhu and Marcotte, 1996) and operator splittings (Davis and Yin, 2017; Vũ, 2013). It can be interpreted as an intermediate notion between monotonicity and strong monotonicity. In general, it is stronger than monotonicity and Lipschitzness of the operator, e.g., simple bilinear games are non-cocoercive. From Cauchy-Swartz's inequality, one can show that a ℓ-co-coercive operator is ℓ-Lipschitz. In single-objective minization, one can prove the converse statement by using convex duality. Thus, a gradient of a function is L–co-coercive if and only if the function is convex and L-smooth (i.e. L-Lipschitz gradients) (Bauschke et al., 2011). However, in general, a L-Lipchitz operator is *not* L–co-coercive. Star-cocoercivity is a new notion recently introduced in Loizou et al. (2021) and is weaker than classical cocoercivity and can be achieved via a proper transformation of quasi-monotone Lipschitz operator (Gorbunov et al., 2022b). Moreover, any μ-quasi strongly monotone L-Lipschitz operator F is ℓ-star-cocoercive with $\ell \in [L, {}^{L^2}/_{\mu}]$ and there exist examples of operators that are quasi-strongly monotone and star-cocoercive but neither monotone nor Lipschitz (Loizou et al., 2021).

Coordinate and zeroth-order methods for solving min-max problems and VIPs. Coordinate methods for solving VIPs are rarely considered in the literature. The most relevant results are given in the literature on zeroth-order methods for solving min-max problems. Although some of them can be easily extended to the coordinate versions of methods for solving VIPs, these methods are usually considered and analyzed for min-max problems. The closest work to our paper is Sadiev et al. (2021): they propose and analyze several zeroth-order variants of SGDA and Stochastic EG with two-point feedback oracle for solving strongly-convex-strongly-concave and convex-concave smooth min-max problems with bounded domain. Moreover, Sadiev et al. (2021) consider firmly smooth convex-concave min-max problems which is an analog of cocoercivity for min-max problems. There are also papers focusing on different problems like non-sonvex-strongly-concave smooth min-max problems (Liu et al., 2020; Wang et al., 2020), non-smooth strongly-convex-strongly-concave and convex-concave min-max problems (Beznosikov et al., 2020b) and on different methods like ones that use one-point feedback oracle (Beznosikov et al., 2021a). These works are less relevant to our paper than Sadiev et al. (2021). Moreover, the results derived in these papers are inferior to the ones from Sadiev et al. (2021).

B MISSING DETAILS ON NUMERICAL EXPERIMENTS

The code for the experiments is available here: https://github.com/hugobb/sgda.

B.1 Setup

We consider the special case of (1) with F and R defined as follows:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x), \quad F_i(x) = \mathbf{A}_i x + b_i, \tag{20}$$

$$R(x) = \lambda\|x\|_1 + \delta_{B_r(0)}(x) = \lambda\|x\|_1 + \begin{cases} 0, & \text{if } \|x\|_\infty \le r, \\ +\infty, & \text{if } \|x\|_\infty > r, \end{cases} \tag{21}$$

where each matrix $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ is non-symmetric with all eigenvalues with strictly positive real part, $b_i \in \mathbb{R}^d$, $r > 0$ is the radius of ℓ_∞-ball, and $\lambda \ge 0$ is regularization parameter. One can show (see Example 6.22 from Beck (2017)) that for the given $R(x)$ prox operator has an explicit formula:

$$\mathrm{prox}_{\gamma R}(x) = \mathrm{sign}\,(x) \min\left\{\max\left\{|x| - \gamma\lambda, 0\right\}, r\right\}, \tag{22}$$

where $\mathrm{sign}(\cdot)$ and $|\cdot|$ are component-wise operators. The considered problem generalizes the following quadratic game:

$$\min_{\|x_1\|_\infty \le r}\max_{\|x_2\|_\infty \le r} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}x_1^\top \mathbf{A}_{1,i}x_1 + x_1^\top \mathbf{A}_{2,i}x_2 - \frac{1}{2}x_2^\top \mathbf{A}_{3,i}x_2 + b_{1,i}^\top x_1 - b_{2,i}^\top x_2 + \lambda\|x_1\|_1 - \lambda\|x_2\|_1$$

with $\mu_i \mathbf{I} \preccurlyeq \mathbf{A}_{1,i} \preccurlyeq L_i\mathbf{I}$ and $\mu_i \mathbf{I} \preccurlyeq \mathbf{A}_{3,i} \preccurlyeq L_i\mathbf{I}$. Indeed, the above problem is a special case of (1)+(21) with

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{1,i} & \mathbf{A}_{2,i} \\ -\mathbf{A}_{2,i} & \mathbf{A}_{3,i} \end{pmatrix}, \quad b_i = \begin{pmatrix} b_{1,i} \\ b_{2,i} \end{pmatrix},$$
$$R(x) = \lambda\|x_1\|_1 + \lambda\|x_2\|_1 + \delta_{B_r(0)}(x_1) + \delta_{B_r(0)}(x_2).$$

In our experiments, to generate the non-symmetric matrices $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ defined in (21), we first sample real random matrices $\mathbf{B}_i$ where the elements of the matrices are sampled from a normal distribution. We then compute the eigendecomposition of the matrices $\mathbf{B}_i = \mathbf{Q}_i\mathbf{D}_i\mathbf{Q}_i^{-1}$, where the $\mathbf{D}_i$ are diagonal matrices with complex numbers on the diagonal. Next, we construct the matrices $\mathbf{A}_i = \Re(\mathbf{Q}_i\mathbf{D}_i^+\mathbf{Q}_i^{-1})$ where $\Re(\mathbf{M})_{i,j} = \Re(\mathbf{M}_{i,j})$ and $\mathbf{D}_i^+$ is obtained by transforming all the elements of $\mathbf{D}_i$ to have positive real part. This process ensures that the eigenvalues of $\mathbf{A}_i$ all have positive real part, and thus that $F(x)$ is strongly monotone and cocoercive. The $b_i \in \mathbb{R}^d$ are sampled from a normal distribution with variance $^{100}/_{d}$. For all the experiments we choose $n = 1000$ and $d = 100$. For the distributed experiments we simulate $m = 10$ nodes on a single machine with 2 CPUs.

B.2 Additional Numerical Experiments with Distributed Methods

In the main part, we reported the numerical results on the comparison of QSGDA, DIANA-SGDA, and VR-DIANA-SGDA applied to solve a distributed version of the quadratic game, in which we assume that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$ with each $\{F_i\}_{i=1}^n$ having similar form to (20). Fig. 3 shows the results for the problem with $R(x) = 0$. In Fig. 4, we present the results for the problem with $R(x)$ defined in (21). The behavior of the methods in this case is very similar to the case without regularization $R(x)$.

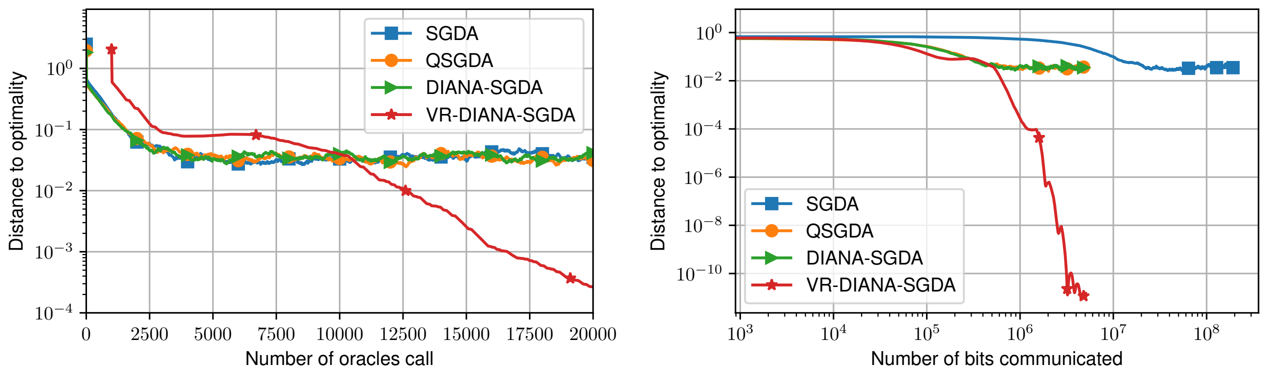

Figure 4: Results on distributed quadratic games with constraints. **Letf:** Number of oracle calls. **Right:** Number of bits communicated between nodes.

However, in both Fig. 3 and 4, DIANA-SGDA performs similarly to QSGDA since the noise σ^2 is larger than the dissimilarity constant ζ_*^2. To illustrate further the difference between DIANA-SGDA and QSGDA, we conduct an additional experiment with full-batched methods ($\sigma = 0$), see Fig. 5. We consider the full-batch version of QSGDA and DIANA-SGDA. This enables us to separate the noise coming from the quantization from the noise coming from the stochasticity. We observe that when using full-batch DIANA-SGDA converges linearly to the solution while QSGDA only converges to a neighborhood of the solution. An interesting observation is that although the convergence is linear, the distance to optimality is not monotonically decreasing, this does not contradicts the theory.

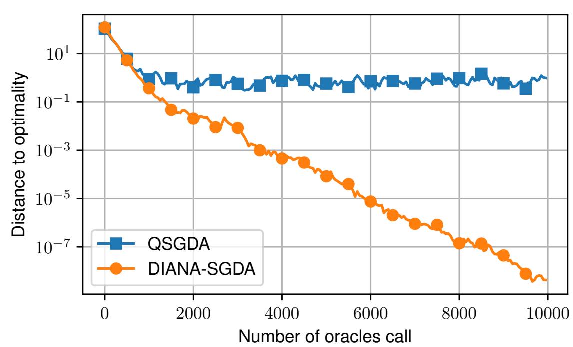

Figure 5: QSGDA vs DIANA-SGDA: DIANA-SGDA converges linearly to the solution while QSGDA only converges to a neighborhood of the solution.

C AUXILIARY RESULTS AND TECHNICAL LEMMAS

Useful inequalities. In our proofs, we often apply the following inequalities that hold for any $a, b \in \mathbb{R}^d$ and $\alpha > 0$:

$$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2, \tag{23}$$

$$\langle a, b\rangle \leq \frac{1}{2\alpha}\|a\|^2 + \frac{\alpha}{2}\|b\|^2. \tag{24}$$

Useful lemmas. The following lemma from Stich (2019) allows us to derive the rates of convergence to the exact solution.

Lemma C.1 (Simplified version of Lemma 3 from Stich (2019))**.** Let the non-negative sequence $\{r_k\}_{k\geq 0}$ satisfy the relation

$$r_{k+1} \leq (1 - a\gamma_k) r_k + c\gamma_k^2$$

for all $k \geq 0$, parameters $a > 0$, $c \geq 0$, and any non-negative sequence $\{\gamma_k\}_{k\geq 0}$ such that $\gamma_k \leq {}^1\!/_h$ for some $h \geq a$, $h > 0$. Then, for any $K \geq 0$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}
\text{if } K \leq \frac{h}{a}, \qquad & \gamma_k = \frac{1}{h},\\
\text{if } K > \frac{h}{a} \text{ and } k < k_0, \qquad & \gamma_k = \frac{1}{h},\\
\text{if } K > \frac{h}{a} \text{ and } k \geq k_0, \qquad & \gamma_k = \frac{2}{a(\kappa + k - k_0)},
\end{aligned}$$

where $\kappa = {}^{2h}\!/_a$ and $k_0 = \lceil {}^K\!/_2 \rceil$. For this choice of γ_k the following inequality holds:

$$r_K \leq \frac{32 h r_0}{a} \exp\left(-\frac{aK}{2h}\right) + \frac{36c}{a^2 K}.$$

In the analysis of monotone case, we rely on the classical result from proximal operators theory.

Lemma C.2 (Theorem 6.39 (iii) from Beck (2017))**.** Let R be a proper lower semicontinuous convex function and $x^+ = \mathrm{prox}_{\gamma R}(x)$. Then for all $z \in \mathbb{R}^d$ the following inequality holds:

$$\langle x^+ - x, z - x^+\rangle \geq \gamma\left(R(x^+) - R(z)\right).$$

Finally, we rely on the following technical lemma for handling the sums arising in the proofs for the monotone case.

Lemma C.3. Let $K > 0$ be a positive integer and $\eta_1, \eta_2, \ldots, \eta_K$ be random vectors such that $\mathbb{E}_k[\eta_k] := \mathbb{E}[\eta_k \mid \eta_1, \ldots, \eta_{k-1}] = 0$ for $k = 2, \ldots, K$. Then

$$\mathbb{E}\left[\left\|\sum_{k=1}^K \eta_k\right\|^2\right] = \sum_{k=1}^K \mathbb{E}[\|\eta_k\|^2]. \tag{25}$$

Proof. We start with the following derivation:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^K \eta_k\right\|^2\right] &= \mathbb{E}[\|\eta_K\|^2] + 2\mathbb{E}\left[\left\langle \eta_K, \sum_{k=1}^{K-1}\eta_k\right\rangle\right] + \mathbb{E}\left[\left\|\sum_{k=1}^{K-1}\eta_k\right\|^2\right]\\
&= \mathbb{E}[\|\eta_K\|^2] + 2\mathbb{E}\left[\mathbb{E}_K\left[\left\langle \eta_K, \sum_{k=1}^{K-1}\eta_k\right\rangle\right]\right] + \mathbb{E}\left[\left\|\sum_{k=1}^{K-1}\eta_k\right\|^2\right]\\
&= \mathbb{E}[\|\eta_K\|^2] + 2\mathbb{E}\left[\left\langle \mathbb{E}_K[\eta_K], \sum_{k=1}^{K-1}\eta_k\right\rangle\right] + \mathbb{E}\left[\left\|\sum_{k=1}^{K-1}\eta_k\right\|^2\right]
\end{aligned}$$

$$= \quad \mathbb{E}[\|\eta_K\|^2] + \mathbb{E}\left[\left\|\sum_{k=1}^{K-1} \eta_k\right\|^2\right].$$

Applying similar steps to $\mathbb{E}\left[\left\|\sum_{k=1}^{K-1} \eta_k\right\|^2\right], \mathbb{E}\left[\left\|\sum_{k=1}^{K-2} \eta_k\right\|^2\right], \ldots, \mathbb{E}\left[\left\|\sum_{k=1}^{2} \eta_k\right\|^2\right]$, we get the result. $\square$

D PROOFS OF THE MAIN RESULTS

In this section, we provide complete proofs of our main results.

D.1 Quasi-Strongly Monotone Case

We start with the case when F satisfies (3) with $\mu > 0$. For readers convenience, we restate the theorems below.

Theorem D.1 (Theorem 2.2)**.** Let F be μ-quasi-strongly monotone with $\mu > 0$ and Assumption 2.1 hold. Assume that

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{2(A+CM)}\right\} \tag{26}$$

for some $M > {}^B/_\rho$. Then for the Lyapunov function $V_k = \|x^k - x^{*,k}\|^2 + M\gamma^2\sigma_k^2$, and for all $k \geq 0$ we have

$$\mathbb{E}[V_k] \quad \leq \quad \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^k \mathbb{E}[V_0] + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - {}^B/_M\right\}}. \tag{27}$$

Proof. First of all, we recall a well-known fact about proximal operators: for any solution x^* of (1) we have

$$x^* = \mathrm{prox}_{\gamma R}(x^* - \gamma F(x^*)). \tag{28}$$

Using this and non-expansiveness of proximal operator, we derive

$$\begin{aligned}
\|x^{k+1} - x^{*,k+1}\|^2 &\leq \|x^{k+1} - x^{*,k}\|^2 \\
&= \left\|\mathrm{prox}_{\gamma R}(x^k - \gamma g^k) - \mathrm{prox}_{\gamma R}(x^{*,k} - \gamma F(x^{*,k}))\right\|^2 \\
&\leq \left\|x^k - \gamma g^k - x^{*,k} - \gamma F(x^{*,k})\right\|^2 \\
&= \|x^k - x^{*,k}\|^2 - 2\gamma\left\langle x^k - x^{*,k}, g^k - F(x^{*,k})\right\rangle + \gamma^2\|g^k - F(x^{*,k})\|^2.
\end{aligned}$$

Next, we take an expectation $\mathbb{E}_k[\cdot]$ w.r.t. the randomness at iteration k and get

$$\begin{aligned}
\mathbb{E}_k\left[\|x^{k+1} - x^{*,k+1}\|^2\right] &= \|x^k - x^{*,k}\|^2 - 2\gamma\left\langle x^k - x^{*,k}, F(x^k) - F(x^{*,k})\right\rangle \\
&\quad + \gamma^2\mathbb{E}_k\left[\left\|g^k - F(x^{*,k})\right\|^2\right] \\
&\overset{(6)}{\leq} \|x^k - x^{*,k}\|^2 - 2\gamma\left\langle x^k - x^*, F(x^k) - F(x^{*,k})\right\rangle \\
&\quad + \gamma^2\left(2A\left\langle x^k - x^{*,k}, F(x^k) - F(x^{*,k})\right\rangle + B\sigma_k^2 + D_1\right).
\end{aligned}$$

Summing up this inequality with (7) multiplied by $M\gamma^2$, we obtain

$$\begin{aligned}
&\mathbb{E}_k\left[\|x^{k+1} - x^{*,k+1}\|^2\right] + M\gamma^2\mathbb{E}_k[\sigma_{k+1}^2] \\
&\quad \leq \|x^k - x^{*,k}\|^2 - 2\gamma\left\langle x^k - x^{*,k}, F(x^k) - F(x^{*,k})\right\rangle \\
&\qquad + \gamma^2\left(2A\left\langle x^k - x^{*,k}, F(x^k) - F(x^{*,k})\right\rangle + B\sigma_k^2 + D_1\right) \\
&\qquad + M\gamma^2\left(2C\left\langle x^k - x^{*,k}, F(x^k) - F(x^{*,k})\right\rangle + (1-\rho)\sigma_k^2 + D_2\right) \\
&\quad = \|x^k - x^{*,k}\|^2 + M\gamma^2\left(1 - \rho + \frac{B}{M}\right)\sigma_k^2 + \gamma^2(D_1 + MD_2) \\
&\qquad - 2\gamma\left(1 - \gamma(A + CM)\right)\left\langle x^k - x^{*,k}, F(x^k) - F(x^{*,k})\right\rangle.
\end{aligned} \tag{29}$$

Since $\gamma \leq \frac{1}{2(A+CM)}$ the factor $-2\gamma\left(1 - \gamma(A+CM)\right)$ is non-positive. Therefore, applying strong quasi-monotonicity of F, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|x^{k+1} - x^{*,k+1}\|^2 + M\gamma^2\sigma_{k+1}^2\right] &\leq \left(1 - 2\gamma\mu\left(1 - \gamma(A+CM)\right)\right)\|x^k - x^{*,k}\|^2 \\
&\quad + M\gamma^2\left(1 - \rho + \frac{B}{M}\right)\sigma_k^2 + \gamma^2(D_1 + MD_2).
\end{aligned}$$

Using $\gamma \leq \frac{1}{2(A+CM)}$ and the definition $V_k = \|x^k - x^{*,k}\|^2 + M\gamma^2\sigma_k^2$, we get

$$\begin{aligned}\mathbb{E}_k\left[V_{k+1}\right] &\leq (1-\gamma\mu)\,\|x^k - x^{*,k}\|^2 + M\gamma^2\left(1-\rho+\frac{B}{M}\right)\sigma_k^2 + \gamma^2(D_1+MD_2)\\ &\leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)V_k + \gamma^2(D_1+MD_2).\end{aligned}$$

Next, we take the full expectation from the above inequality and establish the following recurrence:

$$\mathbb{E}\left[V_{k+1}\right] \leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)\mathbb{E}[V_k] + \gamma^2(D_1+MD_2). \tag{30}$$

Unrolling the recurrence, we derive

$$\begin{aligned}\mathbb{E}\left[V_k\right] &\leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^k\mathbb{E}[V_0] + \gamma^2(D_1+MD_2)\sum_{t=0}^{k-1}\left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^t\\ &\leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^k\mathbb{E}[V_0] + \gamma^2(D_1+MD_2)\sum_{t=0}^{\infty}\left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^t\\ &= \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^k\mathbb{E}[V_0] + \frac{\gamma^2(D_1+MD_2)}{\min\left\{\gamma\mu, \rho - {}^{B}/_{M}\right\}},\end{aligned}$$

which finishes the proof. □

Using this and Lemma C.1, we derive the following result about the convergence to the exact solution.

Corollary D.2 (Corollary 2.3)**.** Let the assumptions of Theorem 2.2 hold. Consider two possible cases.

1. Let $D_1 = D_2 = 0$. Then, for any $K \geq 0$, $M = {}^{2B}/_{\rho}$, and

$$\gamma = \min\left\{\frac{1}{\mu}, \frac{1}{2(A + {}^{2BC}/_{\rho})}\right\} \tag{31}$$

   we have

$$\mathbb{E}[V_K] \leq \mathbb{E}[V_0]\exp\left(-\min\left\{\frac{\mu}{2(A+{}^{2BC}/_{\rho})}, \frac{\rho}{2}\right\}K\right). \tag{32}$$

2. Let $D_1 + MD_2 > 0$. Then, for any $K \geq 0$ and $M = {}^{2B}/_{\rho}$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}&\text{if } K \leq \frac{h}{\mu}, & \gamma_k &= \frac{1}{h},\\ &\text{if } K > \frac{h}{\mu} \text{ and } k < k_0, & \gamma_k &= \frac{1}{h},\\ &\text{if } K > \frac{h}{\mu} \text{ and } k \geq k_0, & \gamma_k &= \frac{2}{\mu(\kappa + k - k_0)},\end{aligned} \tag{33}$$

   where $h = \max\{2(A + {}^{2BC}/_{\rho}), {}^{2\mu}/_{\rho}\}$, $\kappa = {}^{2h}/_{\mu}$ and $k_0 = \lceil {}^{K}/_{2}\rceil$. For this choice of γ_k the following inequality holds:

$$\begin{aligned}\mathbb{E}[V_K] \leq\ & 32\max\left\{\frac{2(A+{}^{2BC}/_{\rho})}{\mu}, \frac{2}{\rho}\right\}\mathbb{E}[V_0]\exp\left(-\min\left\{\frac{\mu}{2(A+{}^{2BC}/_{\rho})}, \frac{\rho}{4}\right\}K\right)\\ &+\frac{36(D_1 + {}^{2BD_2}/_{\rho})}{\mu^2 K}.\end{aligned} \tag{34}$$

Proof. The first part of the corollary follows from Theorem 2.2 due to

$$\left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^K = \left(1-\min\left\{\gamma\mu, \frac{\rho}{2}\right\}\right)^K \leq \exp\left(-\min\left\{\gamma\mu, \frac{\rho}{2}\right\}K\right).$$

Plugging (31) in the above inequality, we derive (32). Next, we consider the case when $D_1 + MD_2 > 0$. First, we notice that (30) holds for non-constant stepsizes γ_k such that

$$0 < \gamma_k \le \min\left\{\frac{1}{\mu}, \frac{1}{2(A+CM)}\right\}.$$

Therefore, for any $k \ge 0$ we have

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\right] &\le \left(1 - \min\left\{\gamma_k\mu, \rho - \frac{B}{M}\right\}\right)\mathbb{E}[V_k] + \gamma_k^2(D_1 + MD_2)\\ &\overset{M=2B/\rho}{=} \left(1 - \min\left\{\gamma_k\mu, \rho/2\right\}\right)\mathbb{E}[V_k] + \gamma_k^2(D_1 + 2BD_2/\rho).\end{aligned}$$

Secondly, we assume that for all $k \ge 0$

$$0 < \gamma_k \le \min\left\{\frac{\rho}{2\mu}, \frac{1}{2(A+CM)}\right\}.$$

Applying this to the recurrence for $\mathbb{E}[V_k]$, we obtain

$$\mathbb{E}\left[V_{k+1}\right] \le (1-\gamma_k\mu)\,\mathbb{E}[V_k] + \gamma_k^2(D_1 + 2BD_2/\rho).$$

It remains to apply Lemma C.1 with $r_k = \mathbb{E}[V_k]$, $a = \mu$, $c = D_1 + 2BD_2/\rho$, and $h = \max\left\{2(A + 2BC/\rho), 2\mu/\rho\right\}$ to the above recurrence. □

D.2 Monotone Case

Next, we consider the case when $\mu = 0$. Before deriving the proof, we provide additional discussion of the setup.

We emphasize that the maximum in (9) is taken over the compact set $\mathcal{C}$ containing the solution set X^*. Therefore, the quantity $\mathrm{Gap}_{\mathcal{C}}(z)$ is a valid measure of convergence (Nesterov, 2007). We point out that the iterates x^k do not have to lie in $\mathcal{C}$. Our analysis works for the problems with unbounded and bounded domains (see Nesterov (2007); Alacaoglu and Malitsky (2021) for similar setups).

Another popular convergence measure for the case when $R(x) \equiv 0$ in (1) is $\|F(x^k)\|^2$. Although the squared norm of the operator is a weaker guarantee, it is easier to compute in practice and better suited for non-monotone problems (Yoon and Ryu, 2021). Nevertheless, $\|F(x^k)\|^2$ is not a valid measure of convergence for (1) with $R(x) \not\equiv 0$. Therefore, we focus on $\mathrm{Gap}_{\mathcal{C}}(z)$ in the monotone case.[7]

Theorem D.3 (Theorem 2.5)**.** Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4 hold. Assume that

$$0 < \gamma \le \frac{1}{2(A + BC/\rho)}. \tag{35}$$

Then for the function $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \ge 0$ we have

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \le\;& \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K}\\ &+\frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + (4A+\ell+8BC/\rho)\cdot\frac{\|x^0-x^{*,0}\|^2}{K}\\ &+\left(4+(4A+\ell+8BC/\rho)\gamma\right)\frac{\gamma B\sigma_0^2}{\rho K}\\ &+\gamma(2+\gamma(4A+\ell+8BC/\rho))(D_1+2BD_2/\rho)\\ &+9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2.\end{aligned} \tag{36}$$

Proof. First, we apply the classical result about proximal operators (Lemma C.2) with $x^+ = x^{k+1}$, $x = x^k - \gamma g^k$, and $z = u$ for arbitrary point $u \in \mathbb{R}^d$:

$$\langle x^{k+1} - x^k + \gamma g^k, u - x^{k+1}\rangle \ge \gamma\left(R(x^{k+1}) - R(u)\right).$$

[7] When $R(x) \equiv 0$, our analysis can be modified to get the guarantees on the squared norm of the operator.

Multiplying by the factor of 2 and making small rearrangement, we get

$$2\gamma\langle g^k, u - x^k\rangle + 2\langle x^{k+1} - x^k, u - x^k\rangle + 2\langle x^{k+1} - x^k + \gamma g^k, x^k - x^{k+1}\rangle \quad\geq\quad 2\gamma\left(R(x^{k+1}) - R(u)\right)$$

implying

$$\begin{aligned} 2\gamma\left(\langle F(x^k), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad&\leq\quad 2\langle x^{k+1} - x^k, u - x^k\rangle + 2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad\quad +2\langle x^{k+1} - x^k, x^k - x^{k+1}\rangle + 2\gamma\langle g^k, x^k - x^{k+1}\rangle. \end{aligned}$$

Next, we use a squared norm decomposition $\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle$, and obtain

$$\begin{aligned} 2\gamma\left(\langle F(x^k), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad&\leq\quad \|x^{k+1} - x^k\|^2 + \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad\quad -2\|x^{k+1} - x^k\|^2 + 2\gamma\langle g^k, x^k - x^{k+1}\rangle. \end{aligned} \tag{37}$$

Then, due to $2\langle a, b\rangle \leq \|a\|^2 + \|b\|^2$ we have

$$\begin{aligned} 2\gamma\left(\langle F(x^k), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad&\leq\quad \|x^{k+1} - x^k\|^2 + \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad\quad -2\|x^{k+1} - x^k\|^2 + \gamma^2\|g^k\|^2 + \|x^k - x^{k+1}\|^2 \\ &=\quad \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma\langle F(x^k) - g^k, x^k - u\rangle + \gamma^2\|g^k\|^2. \end{aligned}$$

Monotonicity of F implies $\langle F(u), x^k - u\rangle \leq \langle F(x^k), x^k - u\rangle$, allowing us to continue our derivation as follows:

$$\begin{aligned} 2\gamma\left(\langle F(u), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad&\leq\quad \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma\langle F(x^k) - g^k, x^k - u\rangle + \gamma^2\|g^k\|^2 \\ &=\quad \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad\quad +\gamma^2\|g^k - g^{*,k} + g^{*,k}\|^2 \\ &\overset{(23)}{\leq}\quad \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad\quad +2\gamma^2\|g^k - g^{*,k}\|^2 + 2\gamma^2\|g^{*,k}\|^2. \end{aligned}$$

Summing up the above inequality for $k = 0, 1, \ldots, K-1$, we get

$$\begin{aligned} 2\gamma\sum_{k=0}^{K-1}\left(\langle F(u), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad&\leq\quad \sum_{k=0}^{K-1}\|x^k - u\|^2 - \sum_{k=0}^{K-1}\|x^{k+1} - u\|^2 \\ &\quad\quad +2\gamma^2\sum_{k=0}^{K-1}\|g^{*,k}\|^2 \\ &\quad\quad +2\gamma\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad\quad +2\gamma^2\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2 \\ &=\quad \|x^0 - u\|^2 - \|x^K - u\|^2 + 2\gamma^2\sum_{k=0}^{K-1}\|g^{*,k}\|^2 \\ &\quad\quad +2\gamma\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle \end{aligned}$$

$$+2\gamma^2\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2.$$

Next, we divide both sides by $2\gamma K$

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\left(\langle F(u), x^k - u\rangle + R(x^{k+1}) - R(u)\right) &\leq \frac{\|x^0-u\|^2 - \|x^K - u\|^2}{2\gamma K} + \frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^{*,k}\|^2 \\ &+\frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle \\ &+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2\end{aligned}$$

and, after small rearrangement, we obtain

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\left(\langle F(u), x^{k+1} - u\rangle + R(x^{k+1}) - R(u)\right) &\leq \frac{\|x^0-u\|^2 - \|x^K - u\|^2}{2\gamma K} + \frac{\langle F(u), x^K - x^0\rangle}{K} \\ &+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^{*,k}\|^2 \\ &+\frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle \\ &+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2.\end{aligned}$$

Applying Jensen's inequality for convex function R, we get $R\left(\frac{1}{K}\sum_{k=0}^{K-1}x^{k+1}\right) \leq \frac{1}{K}\sum_{k=0}^{K-1}R(x^{k+1})$. Plugging this in the previous inequality, we derive for u^* being a projection of u on X^*

$$\begin{aligned}&\left\langle F(u), \left(\frac{1}{K}\sum_{k=0}^{K-1}x^{k+1}\right) - u\right\rangle + R\left(\frac{1}{K}\sum_{k=0}^{K-1}x^{k+1}\right) - R(u) \\ &\quad\leq \frac{\|x^0-u\|^2 - \|x^K - u\|^2}{2\gamma K} + \frac{\langle F(u), x^K - x^0\rangle}{K} + \frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^{*,k}\|^2 \\ &\quad+\frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle + \frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2 \\ &\quad\overset{(24)}{\leq} \frac{\|x^0-u\|^2 - \|x^K - u\|^2}{2\gamma K} + \frac{\|x^K - x^0\|^2}{4\gamma K} + \frac{4\gamma}{K}\|F(u) - F(u^*) + F(u^*)\|^2 \\ &\quad+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^{*,k}\|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle + \frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2 \\ &\quad\overset{(23)}{\leq} \frac{\|x^0-u\|^2 - \|x^K - u\|^2}{2\gamma K} + \frac{\|x^0 - u\|^2 + \|x^K - u\|^2}{2\gamma K} + \frac{8\gamma}{K}\|F(u) - F(u^*)\|^2 \\ &\quad+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^{*,k}\|^2 + 8\gamma\|F(u^*)\|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle \\ &\quad+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - g^{*,k}\|^2 \\ &\quad\overset{(4)}{\leq} \frac{\|x^0-u\|^2}{\gamma K} + \frac{8\gamma\ell^2\|u-u^*\|^2}{K} + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2\end{aligned}$$

$$+\frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^k-u\rangle+\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k-g^{*,k}\|^2.$$

Next, we take maximum from the both sides in $u\in\mathcal{C}$, which gives $\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)$ in the left-hand side by definition (9), and take the expectation of the result:

$$\begin{aligned}
\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{\mathbb{E}\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{\gamma K}+\frac{8\gamma\ell^2\mathbb{E}\left[\max_{u\in\mathcal{C}}\|u-u^*\|^2\right]}{K}\\
&\quad+9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2\\
&\quad+\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^k-u\rangle\right]+\frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k-g^{*,k}\|^2\right]\\
&\leq \frac{\mathbb{E}\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{\gamma K}+\frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K}+9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2\\
&\quad+\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^k-u\rangle\right]\\
&\quad+\frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k-g^{*,k}\|^2\right].
\end{aligned}\tag{38}$$

In the last step, we also use that $X^*\subset\mathcal{C}$ and $\Omega_{\mathcal{C}}:=\max_{x,y\in\mathcal{C}}\|x-y\|$ (Assumption 2.4).

It remains to upper bound the terms from the last two lines of (38). We start with the first one. Since

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^k\rangle\right] &= \mathbb{E}\left[\sum_{k=0}^{K-1}\langle\mathbb{E}[F(x^k)-g^k\mid x^k],x^k\rangle\right]=0,\\
\mathbb{E}\left[\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^0\rangle\right] &= \sum_{k=0}^{K-1}\left\langle\mathbb{E}[F(x^k)-g^k],x^0\right\rangle=0,
\end{aligned}$$

we have

$$\begin{aligned}
\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^k-u\rangle\right] &= \frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^k\rangle\right]\\
&\qquad+\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,-u\rangle\right]\\
&= \frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,-u\rangle\right]\\
&= \frac{1}{K}\mathbb{E}\left[\sum_{k=0}^{K-1}\langle F(x^k)-g^k,x^0\rangle\right]\\
&\qquad+\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k,-u\rangle\right]\\
&= \mathbb{E}\left[\max_{u\in\mathcal{C}}\left\langle\frac{1}{K}\sum_{k=0}^{K-1}(F(x^k)-g^k),x^0-u\right\rangle\right]\\
&\overset{(24)}{\leq} \mathbb{E}\left[\max_{u\in\mathcal{C}}\left\{\frac{\gamma K}{2}\left\|\frac{1}{K}\sum_{k=0}^{K-1}(F(x^k)-g^k)\right\|^2+\frac{1}{2\gamma K}\|x^0-u\|^2\right\}\right]\\
&= \frac{\gamma}{2K}\mathbb{E}\left[\left\|\sum_{k=0}^{K-1}(F(x^k)-g^k)\right\|^2\right]+\frac{1}{2\gamma K}\max_{u\in\mathcal{C}}\|x^0-u\|^2.
\end{aligned}$$

We notice that $\mathbb{E}[F(x^k) - g^k \mid F(x^0) - g^0, \ldots, F(x^{k-1}) - g^{k-1}] = 0$ for all $k \geq 1$, i.e., conditions of Lemma C.3 are satisfied. Therefore, applying Lemma C.3, we get

$$\begin{aligned}\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle\right] &\leq \frac{\gamma}{2K}\sum_{k=0}^{K-1}\mathbb{E}[\|F(x^k) - g^k\|^2] \\ &\quad + \frac{1}{2\gamma K}\max_{u\in\mathcal{C}}\|x^0 - u\|^2. \end{aligned} \tag{39}$$

Combining (38) and (39), we derive

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2 \\ &\quad + \frac{\gamma}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k - F(x^k)\|^2\right] \\ &\quad + \frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k - g^{*,k}\|^2\right] \\ &\overset{(23)}{\leq} \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2 \\ &\quad + \frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|F(x^k) - g^{*,k}\|^2\right] + \frac{2\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k - g^{*,k}\|^2\right]. \end{aligned} \tag{40}$$

Using ℓ-star-cocoercivity of F together with the first part of Assumption 2.1, we continue our derivation as follows:

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + 2\gamma D_1 + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2 \\ &\quad + \frac{\gamma(4A+\ell)}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle\right] + \frac{2\gamma B}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2\right] \\ &= \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + 2\gamma D_1 + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2 \\ &\quad + \frac{\gamma(4A+\ell)}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle\right] \\ &\quad + \frac{2\gamma B}{K}\left(1 + \frac{1}{\rho}\right)\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2\right] - \frac{2\gamma B}{\rho K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2\right]. \end{aligned}$$

Next, we use the second part of Assumption 2.1 and get

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + 2\gamma D_1 + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2 \\ &\quad + \frac{\gamma(4A+\ell)}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle F(x^k) - g^{*,k}, x^k - x^{*,k}\rangle\right] \\ &\quad + \frac{2\gamma B}{K}\left(1 + \frac{1}{\rho}\right)\sum_{k=1}^{K-1}\mathbb{E}\left[2C\langle F(x^{k-1}) - g^{*,k-1}, x^{k-1} - x^{*,k-1}\rangle\right] \\ &\quad + \frac{2\gamma B}{K}\left(1 + \frac{1}{\rho}\right)\sum_{k=1}^{K-1}\mathbb{E}\left[(1-\rho)\sigma_{k-1}^2 + D_2\right] \\ &\quad + \frac{2\gamma B}{K}\left(1 + \frac{1}{\rho}\right)\sigma_0^2 - \frac{2\gamma B}{\rho K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K}+\frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K}+\frac{2\gamma B(1+1/\rho)}{K}\sigma_0^2\\
&\quad +2\gamma\left(D_1+B(1+1/\rho)D_2\right)\\
&\quad +9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2+\frac{\gamma(4A+\ell)}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle F(x^k)-g^{*,k},x^k-x^{*,k}\rangle\right]\\
&\quad +\frac{2\gamma B}{K}\left(1+\frac{1}{\rho}\right)\sum_{k=0}^{K-2}\mathbb{E}\left[2C\langle F(x^k)-g^{*,k},x^k-x^{*,k}\rangle+(1-\rho)\sigma_k^2\right]\\
&\quad -\frac{2\gamma B}{\rho K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2\right]\\
&\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K}+\frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K}+\frac{2\gamma B(1+1/\rho)}{K}\sigma_0^2\\
&\quad +2\gamma\left(D_1+B(1+1/\rho)D_2\right)+9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2\\
&\quad +\left(4A+\ell+4BC(1+1/\rho)\right)\frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle F(x^k)-g^{*,k},x^k-x^{*,k}\rangle\right]\\
&\quad +\frac{2\gamma B}{K}(1-\rho)\left(1+\frac{1}{\rho}\right)\sum_{k=0}^{K-2}\mathbb{E}\left[\sigma_k^2\right]-\frac{2\gamma B}{\rho K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2\right].
\end{aligned}
$$

Since $(1-\rho)\left(1+1/\rho\right)=-\rho+1/\rho\leq 1/\rho$, the last row is non-positive and we have

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K}+\frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K}+\frac{2\gamma B(1+1/\rho)}{K}\sigma_0^2\\
&\quad +2\gamma\left(D_1+B(1+1/\rho)D_2\right)+9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2\\
&\quad +\frac{\gamma\left(4A+\ell+4BC(1+1/\rho)\right)}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\langle F(x^k)-g^{*,k},x^k-x^{*,k}\rangle\right].
\end{aligned}
\tag{41}
$$

Note that inequality (29) from the proof of Theorem 2.2 is derived using Assumption 2.1 only. With $M=B/\rho$ it gives

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1}-x^{*,k+1}\|^2\right]+\frac{\gamma^2B}{\rho}\mathbb{E}[\sigma_{k+1}^2] &\leq \mathbb{E}\left[\|x^k-x^{*,k}\|^2\right]+\frac{\gamma^2B}{\rho}\mathbb{E}\left[\sigma_k^2\right]+\gamma^2(D_1+BD_2/\rho)\\
&\quad -2\gamma\left(1-\gamma(A+BC/\rho)\right)\mathbb{E}\left[\left\langle x^k-x^{*,k},F(x^k)-g^{*,k}\right\rangle\right].
\end{aligned}
$$

Since $\gamma\leq 1/2(A+BC/\rho)$ we obtain

$$
\begin{aligned}
\gamma\mathbb{E}\left[\left\langle x^k-x^{*,k},F(x^k)-g^{*,k}\right\rangle\right] &\leq \mathbb{E}\left[\|x^k-x^{*,k}\|^2\right]+\frac{\gamma^2B}{\rho}\mathbb{E}\left[\sigma_k^2\right]-\mathbb{E}\left[\|x^{k+1}-x^{*,k+1}\|^2\right]\\
&\quad -\frac{\gamma^2B}{\rho}\mathbb{E}[\sigma_{k+1}^2]+\gamma^2(D_1+BD_2/\rho).
\end{aligned}
$$

Plugging this inequality in (41), we derive

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K}+\frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K}+\frac{2\gamma B(1+1/\rho)}{K}\sigma_0^2\\
&\quad +2\gamma\left(D_1+B(1+1/\rho)D_2\right)+9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2\\
&\quad +\left(4A+\ell+4BC(1+1/\rho)\right)\cdot\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|x^k-x^{*,k}\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\quad - (4A + \ell + 4BC(1 + 1/\rho)) \cdot \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|x^{k+1} - x^{*,k+1}\|^2\right] \\
&\quad + (4A + \ell + 4BC(1 + 1/\rho)) \cdot \frac{\gamma^2 B}{\rho K}\sum_{k=0}^{K-1}\mathbb{E}\left[\sigma_k^2 - \sigma_{k+1}^2\right] \\
&\quad +\gamma^2\left(4A + \ell + 4BC(1 + 1/\rho)\right) \cdot (D_1 + BD_2/\rho) \\
\leq &\quad \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + (4A + \ell + 8BC/\rho)\cdot\frac{\|x^0 - x^{*,0}\|^2}{K} \\
&\quad + \left(4 + (4A + \ell + 8BC/\rho)\gamma\right)\frac{\gamma B\sigma_0^2}{\rho K} \qquad (42)\\
&\quad +\gamma\left((2 + \gamma(4A + \ell + 8BC/\rho))(D_1 + 2BD_2/\rho) + 9\max_{x^*\in X^*}\|F(x^*)\|^2\right),
\end{aligned}
$$

where in the last inequality we use $1 + 1/\rho \leq 2/\rho$. $\square$

Corollary D.4. Let the assumptions of Theorem 2.5 hold. Then, for all K one can choose γ as

$$
\gamma = \min\left\{\frac{1}{4A + \ell + 8BC/\rho}, \frac{\Omega_{0,\mathcal{C}}\sqrt{\rho}}{\widehat{\sigma}_0\sqrt{B}}, \frac{\Omega_{0,\mathcal{C}}}{\sqrt{K(D_1 + 2BD_2/\rho)}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}, \qquad (43)
$$

where $\Omega_0 := \|x^0 - x^{*,0}\|^2$ and $\Omega_{0,\mathcal{C}}$, $\widehat{\sigma}_0$, and G_* are some upper bounds for $\max_{u\in\mathcal{C}}\|x^0 - u\|$, σ_0, and $\max_{x^*\in X^*}\|F(x^*)\|$ respectively. This choice of γ implies $\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right]$ equals

$$
\mathcal{O}\left(\frac{(A + \ell + BC/\rho)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}\widehat{\sigma}_0\sqrt{B}}{\sqrt{\rho}K} + \frac{\Omega_{0,\mathcal{C}}(\sqrt{D_1 + BD_2/\rho} + G_*)}{\sqrt{K}}\right).
$$

Proof. First of all, the choice of γ from (43) implies (35) since

$$
\frac{1}{4A + \ell + 8BC/\rho} \leq \frac{1}{2(A + BC/\rho)}.
$$

Using (10), the definitions of $\Omega_{0,\mathcal{C}}$, $\widehat{\sigma}_0$, G_*, and $\gamma \leq 1/(4A+\ell+8BC/\rho)$, we get

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \leq &\quad \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + (4A + \ell + 8BC/\rho)\cdot\frac{\|x^0 - x^{*,0}\|^2}{K} \\
&\quad + \left(4 + (4A + \ell + 8BC/\rho)\gamma\right)\frac{\gamma B\sigma_0^2}{\rho K} \\
&\quad +\gamma\left((2 + \gamma(4A + \ell + 8BC/\rho))(D_1 + 2BD_2/\rho) + 9\max_{x^*\in X^*}\|F(x^*)\|^2\right) \\
\leq &\quad \frac{3\Omega_{0,\mathcal{C}}^2}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \frac{(4A + \ell + 8BC/\rho)\Omega_0^2}{K} \\
&\quad + \left(4 + (4A + \ell + 8BC/\rho)\gamma\right)\frac{\gamma B\widehat{\sigma}_0^2}{\rho K} \\
&\quad +\gamma\left((2 + \gamma(4A + \ell + 8BC/\rho))(D_1 + 2BD_2/\rho) + 9G_*^2\right) \\
\leq &\quad \frac{3\Omega_{0,\mathcal{C}}^2}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \frac{(4A + \ell + 8BC/\rho)\Omega_0^2}{K} + \frac{5\gamma B\widehat{\sigma}_0^2}{\rho K} \\
&\quad +3\gamma\left(D_1 + \frac{2BD_2}{\rho} + 3G_*^2\right).
\end{aligned}
$$

Finally, we apply (43):

$$
\begin{aligned}
\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] &\leq \frac{3\Omega_{0,\mathcal{C}}^2}{2\min\left\{\frac{1}{4A+\ell+{}^{8BC}/\rho}, \frac{\Omega_{0,\mathcal{C}}\sqrt{\rho}}{\widehat{\sigma}_0\sqrt{B}}, \frac{\Omega_{0,\mathcal{C}}}{\sqrt{K(D_1+{}^{2BD_2}/\rho)}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}K} + \frac{1}{\ell}\cdot\frac{8\ell^2\Omega_{\mathcal{C}}^2}{K} \\
&\quad + \frac{(4A+\ell+{}^{8BC}/\rho)\Omega_0^2}{K} + \frac{\Omega_{0,\mathcal{C}}\sqrt{\rho}}{\widehat{\sigma}_0\sqrt{B}}\cdot\frac{\gamma B\widehat{\sigma}_0^2}{\rho K} \\
&\quad + \frac{\Omega_{0,\mathcal{C}}}{\sqrt{K(D_1+{}^{2BD_2}/\rho)}}\cdot 3\left(D_1+\frac{2BD_2}{\rho}\right) + \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\cdot 9G_*^2 \\
&= \mathcal{O}\Bigg(\frac{(A+\ell+{}^{BC}/\rho)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)+\ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}\widehat{\sigma}_0\sqrt{B}}{\sqrt{\rho}K} \\
&\qquad\qquad + \frac{\Omega_{0,\mathcal{C}}(\sqrt{D_1+{}^{BD_2}/\rho}+G_*)}{\sqrt{K}}\Bigg).
\end{aligned}
$$

□

D.3 Cocoercive Case

The upper bound from Theorem 2.5 contains the term proportional to $\max_{x^* \in X^*} \|F(x^*)\|^2$, which is non-zero in general. Therefore, even when there is no noise the method with constant stepsize converges only to some error proportional to $\max_{x^* \in X^*} \|F(x^*)\|^2$. To resolve this issue we assume ℓ-cocoercivity of F, i.e., we assume that

$$\|F(x) - F(y)\|^2 \le \ell \langle F(x) - F(y), x - y \rangle \quad \forall x, y \in \mathbb{R}^d.$$

Theorem D.5 (Theorem 2.6)**.** Let F be ℓ-cocoercive and Assumptions 2.1, 2.4 hold. Assume that

$$0 < \gamma \le \min\left\{\frac{1}{\ell}, \frac{1}{2(A + {}^{BC}/_{\rho})}\right\}. \tag{44}$$

Then for the function $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \ge 0$ we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \quad \le \quad & \frac{3\max_{u\in\mathcal{C}} \|x^0 - u\|^2}{2\gamma K} + (6A + 3\ell + {}^{12BC}/_{\rho}) \cdot \frac{\|x^0 - x^{*,0}\|^2}{K} \\
& + (6 + (6A + 3\ell + {}^{12BC}/_{\rho})\,\gamma)\,\frac{\gamma B \sigma_0^2}{\rho K} \\
& + \gamma(3 + \gamma\,(6A + 3\ell + {}^{12BC}/_{\rho}))(D_1 + {}^{2BD_2}/_{\rho}).
\end{aligned} \tag{45}$$

Proof. We start the proof from (37).

$$\begin{aligned}
2\gamma\left(\langle F(x^k), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad \le \quad & \|x^{k+1} - x^k\|^2 + \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\
& +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\
& -2\|x^{k+1} - x^k\|^2 + 2\gamma\langle g^k, x^k - x^{k+1}\rangle \\
= \quad & \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\
& +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\
& -\|x^{k+1} - x^k\|^2 + 2\gamma\langle F(u), x^k - x^{k+1}\rangle \\
& +2\gamma\langle g^k - F(u), x^k - x^{k+1}\rangle.
\end{aligned}$$

Then, due to $2\langle a, b\rangle \le \|a\|^2 + \|b\|^2$ we have

$$\begin{aligned}
2\gamma\left(\langle F(x^k), x^k - u\rangle + R(x^{k+1}) - R(u)\right) \quad \le \quad & \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\
& +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\
& -\|x^{k+1} - x^k\|^2 + 2\gamma\langle F(u), x^k - x^{k+1}\rangle \\
& +\gamma^2\|g^k - F(u)\|^2 + \|x^k - x^{k+1}\|^2 \\
= \quad & \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\
& +2\gamma\langle F(x^k) - g^k, x^k - u\rangle \\
& +2\gamma\langle F(u), x^k - x^{k+1}\rangle + \gamma^2\|g^k - F(u)\|^2.
\end{aligned}$$

Next, we add $2\gamma\left(\langle F(u), x^{k+1} - u\rangle - \langle F(x^k), x^k - u\rangle\right)$ to both sides of the previous inequality.

$$\begin{aligned}
2\gamma\left(\langle F(u), x^{k+1} - u\rangle + R(x^{k+1}) - R(u)\right) \quad \le \quad & \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\
& +2\gamma\langle F(u) - g^k, x^k - u\rangle + \gamma^2\|g^k - F(u)\|^2 \\
\le \quad & \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \\
& -2\gamma\langle F(x^k) - F(u), x^k - u\rangle \\
& -2\gamma\langle g^k - F(x^k), x^k - u\rangle \\
& +2\gamma^2\|g^k - F(x^k)\|^2 + 2\gamma^2\|F(x^k) - F(u)\|^2.
\end{aligned}$$

Using that F is ℓ-co-cocoercive, we get

$$2\gamma\left(\langle F(u), x^{k+1} - u\rangle + R(x^{k+1}) - R(u)\right) \quad \le \quad \|x^k - u\|^2 - \|x^{k+1} - u\|^2$$

$$
\begin{aligned}
& -\frac{2\gamma}{\ell}\|F(x^k)-F(u)\|^2 \\
& -2\gamma\langle g^k - F(x^k), x^k - u\rangle \\
& +2\gamma^2\|g^k - F(x^k)\|^2 + 2\gamma^2\|F(x^k)-F(u)\|^2 \\
= \;& \|x^k - u\|^2 - \|x^{k+1}-u\|^2 \\
& -\frac{2\gamma}{\ell}\left(1-\gamma\ell\right)\|F(x^k)-F(u)\|^2 \\
& -2\gamma\langle g^k - F(x^k), x^k - u\rangle \\
& +2\gamma^2\|g^k - F(x^k)\|^2.
\end{aligned}
$$

With $\gamma \leq \frac{1}{\ell}$, we have

$$
\begin{aligned}
2\gamma\left(\langle F(u), x^{k+1}-u\rangle + R(x^{k+1}) - R(u)\right) \;\leq\; & \|x^k-u\|^2 - \|x^{k+1}-u\|^2 \\
& -2\gamma\langle g^k - F(x^k), x^k - u\rangle \\
& +2\gamma^2\|g^k - F(x^k)\|^2.
\end{aligned}
$$

Summing up the above inequality for $k = 0, 1, \ldots, K-1$, we get

$$
\begin{aligned}
2\gamma\sum_{k=0}^{K-1}\left(\langle F(u), x^{k+1}-u\rangle + R(x^{k+1}) - R(u)\right) \;\leq\; & \sum_{k=0}^{K-1}\|x^k-u\|^2 - \sum_{k=0}^{K-1}\|x^{k+1}-u\|^2 \\
& +2\gamma^2\sum_{k=0}^{K-1}\|g^k - F(x^k)\|^2 \\
& +2\gamma\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle.
\end{aligned}
$$

Next, we divide both sides by $2\gamma K$

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\left(\langle F(u), x^{k+1}-u\rangle + R(x^{k+1}) - R(u)\right) \;\leq\; & \frac{\|x^0-u\|^2 - \|x^K-u\|^2}{2\gamma K} \\
& +\frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - F(x^k)\|^2 \\
& +\frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle.
\end{aligned}
$$

Applying Jensen's inequality for convex function R, we get $R\left(\frac{1}{K}\sum_{k=0}^{K-1} x^{k+1}\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} R(x^{k+1})$.

$$
\begin{aligned}
& \left\langle F(u), \left(\frac{1}{K}\sum_{k=0}^{K-1} x^{k+1}\right) - u\right\rangle + R\left(\frac{1}{K}\sum_{k=0}^{K-1} x^{k+1}\right) - R(u) \\
& \qquad \leq \frac{\|x^0-u\|^2 - \|x^K-u\|^2}{2\gamma K} \\
& \qquad\quad + \frac{\gamma}{K}\sum_{k=0}^{K-1}\|g^k - F(x^k)\|^2 \\
& \qquad\quad + \frac{1}{K}\sum_{k=0}^{K-1}\langle F(x^k) - g^k, x^k - u\rangle.
\end{aligned}
$$

Next, we take maximum from the both sides in $u \in \mathcal{C}$, which gives $\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)$ in the left-hand side by definition (9), and take the expectation of the result:

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \;\leq\; \frac{\mathbb{E}\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{\gamma K} + \frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k - F(x^k)\|^2\right]
$$

$$+\frac{1}{K}\mathbb{E}\left[\max_{u\in\mathcal{C}}\sum_{k=0}^{K-1}\langle F(x^k)-g^k, x^k-u\rangle\right].$$

Using the estimate (39), we get

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{\max_{u\in\mathcal{C}}\|x^0-u\|^2}{\gamma K}+\frac{\gamma}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k-F(x^k)\|^2\right]\\ &\quad+\frac{\gamma}{2K}\sum_{k=0}^{K-1}\mathbb{E}[\|F(x^k)-g^k\|^2]+\frac{1}{2\gamma K}\max_{u\in\mathcal{C}}\|x^0-u\|^2\\ &\leq \frac{3\max_{u\in\mathcal{C}}\|x^0-u\|^2}{2\gamma K}+\frac{3\gamma}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|g^k-F(x^k)\|^2\right].\end{aligned}$$

It remains to estimate $\frac{1}{K}\sum\limits_{k=0}^{K-1}\mathbb{E}\left[\|g^k-F(x^k)\|^2\right]$. This was done in the previous proof (see from (40) to (42)). Then, we finally have

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\max_{u\in\mathcal{C}}\|x^0-u\|^2}{2\gamma K}+\left(6A+3\ell+{}^{12BC}/_{\rho}\right)\cdot\frac{\|x^0-x^{*,0}\|^2}{K}\\ &\quad+\left(6+\left(6A+3\ell+{}^{12BC}/_{\rho}\right)\gamma\right)\frac{\gamma B\sigma_0^2}{\rho K}\\ &\quad+\gamma(3+\gamma\left(6A+3\ell+{}^{12BC}/_{\rho}\right))(D_1+{}^{2BD_2}/_{\rho}).\end{aligned}$$

□

Corollary D.6. Let the assumptions of Theorem D.5 hold. Then, for all K one can choose γ as

$$\gamma=\min\left\{\frac{1}{6A+3\ell+{}^{12BC}/_{\rho}},\frac{\Omega_{0,\mathcal{C}}\sqrt{\rho}}{\widehat{\sigma}_0\sqrt{B}},\frac{\Omega_{0,\mathcal{C}}}{\sqrt{K(D_1+{}^{2BD_2}/_{\rho})}}\right\},\tag{46}$$

where $\Omega_0:=\|x^0-x^{*,0}\|^2$ and $\Omega_{0,\mathcal{C}}$, and $\widehat{\sigma}_0$ are some upper bounds for $\max_{u\in\mathcal{C}}\|x^0-u\|$, and σ_0 respectively. This choice of γ implies $\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right]$ equals

$$\mathcal{O}\left(\frac{(A+\ell+{}^{BC}/_{\rho})(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)}{K}+\frac{\Omega_{0,\mathcal{C}}\widehat{\sigma}_0\sqrt{B}}{\sqrt{\rho}K}+\frac{\Omega_{0,\mathcal{C}}\sqrt{D_1+{}^{BD_2}/_{\rho}}}{\sqrt{K}}\right).$$

Proof. First of all, the choice of γ from (46) implies (35) since

$$\frac{1}{6A+3\ell+{}^{12BC}/_{\rho}}\leq\frac{1}{2\left(A+{}^{BC}/_{\rho}\right)}.$$

Using (45), the definitions of $\Omega_{0,\mathcal{C}}$, $\widehat{\sigma}_0$, and $\gamma\leq{}^{1}/_{(6A+3\ell+{}^{12BC}/_{\rho})}$, we get

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\leq \frac{3\max_{u\in\mathcal{C}}\|x^0-u\|^2}{2\gamma K}+\left(6A+3\ell+{}^{12BC}/_{\rho}\right)\cdot\frac{\|x^0-x^{*,0}\|^2}{K}\\ &\quad+\left(6+\left(6A+3\ell+{}^{12BC}/_{\rho}\right)\gamma\right)\frac{\gamma B\sigma_0^2}{\rho K}\\ &\quad+\gamma(3+\gamma\left(6A+3\ell+{}^{12BC}/_{\rho}\right))(D_1+{}^{2BD_2}/_{\rho})\\ &\leq \frac{3\Omega_{0,\mathcal{C}}^2}{2\gamma K}+\frac{(6A+3\ell+{}^{12BC}/_{\rho})\Omega_0^2}{K}\\ &\quad+\left(6+\left(6A+3\ell+{}^{12BC}/_{\rho}\right)\gamma\right)\frac{\gamma B\widehat{\sigma}_0^2}{\rho K}\end{aligned}$$

$$
\begin{aligned}
&+\gamma(3+\gamma\left(6A+3\ell+{}^{12BC}/_{\rho}\right))(D_1+{}^{2BD_2}/_{\rho}) \\
\leq\quad & \frac{3\Omega_{0,\mathcal{C}}^2}{2\gamma K}+\frac{(6A+3\ell+{}^{12BC}/_{\rho})\Omega_0^2}{K}+\frac{7\gamma B\widehat{\sigma}_0^2}{\rho K}+4\gamma\left(D_1+\frac{2BD_2}{\rho}\right).
\end{aligned}
$$

Finally, we apply (43):

$$
\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \quad\leq\quad & \frac{3\Omega_{0,\mathcal{C}}^2}{2\min\left\{\frac{1}{6A+3\ell+{}^{12BC}/_{\rho}},\frac{\Omega_{0,\mathcal{C}}\sqrt{\rho}}{\widehat{\sigma}_0\sqrt{B}},\frac{\Omega_{0,\mathcal{C}}}{\sqrt{K(D_1+{}^{2BD_2}/_{\rho})}}\right\}K} \\
&+\frac{(6A+3\ell+{}^{12BC}/_{\rho})\Omega_0^2}{K}+\frac{\Omega_{0,\mathcal{C}}\sqrt{\rho}}{\widehat{\sigma}_0\sqrt{B}}\cdot\frac{\gamma B\widehat{\sigma}_0^2}{\rho K} \\
&+\frac{\Omega_{0,\mathcal{C}}}{\sqrt{K(D_1+{}^{2BD_2}/_{\rho})}}\cdot 4\left(D_1+\frac{2BD_2}{\rho}\right) \\
=\quad & \mathcal{O}\Bigg(\frac{(A+\ell+{}^{BC}/_{\rho})(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)}{K}+\frac{\Omega_{0,\mathcal{C}}\widehat{\sigma}_0\sqrt{B}}{\sqrt{\rho}K} \\
&\qquad+\frac{\Omega_{0,\mathcal{C}}\sqrt{D_1+{}^{BD_2}/_{\rho}}}{\sqrt{K}}\Bigg).
\end{aligned}
$$

$\square$

E SGDA WITH ARBITRARY SAMPLING: MISSING PROOFS AND DETAILS

Algorithm 1 SGDA-AS: Stochastic Gradient Descent-Ascent with Arvitrary Sampling

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, distribution $\mathcal{D}$, stepsize $\gamma > 0$, number of steps K
2: **for** $k = 0$ **to** $K-1$ **do**
3: Sample $\xi^k \sim \mathcal{D}$ independently from previous iterations and compute $g^k = F_{\xi^k}(x^k)$
4: $x^{k+1} = \text{prox}_{\gamma R}(x^k - \gamma g^k)$
5:
6: **end for**

E.1 Proof of Proposition 3.2

Proposition E.1 (Proposition 3.2)**.** Let Assumption 3.1 hold. Then, SGDA satisfies Assumption 2.1 with

$$A = \ell_{\mathcal{D}}, \quad B = 0, \quad \sigma_k^2 \equiv 0, \quad D_1 = 2\sigma_*^2 := 2\max_{x^* \in X^*} \mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x^*) - F(x^*)\|^2\right],$$
$$C = 0, \quad \rho = 1, \quad D_2 = 0.$$

Proof. To prove the result, it is sufficient to derive an upper bound for $\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right]$:

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &= \mathbb{E}_{\mathcal{D}}\left[\|F_{\xi^k}(x^k) - F(x^{*,k})\|^2\right] \\
&\leq 2\mathbb{E}_{\mathcal{D}}\left[\|F_{\xi^k}(x^k) - F_{\xi^k}(x^{*,k})\|^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[\|F_{\xi^k}(x^{*,k}) - F(x^{*,k})\|^2\right] \\
&\overset{Ass.(3.1)}{\leq} 2\ell_{\mathcal{D}}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + 2\sigma_*^2,
\end{aligned}$$

where $\sigma_*^2 := \max_{x^* \in X^*} \mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x^*) - F(x^*)\|^2\right]$. The above inequality implies that Assumption 2.1 holds with

$$A = \ell_{\mathcal{D}}, \quad B = 0, \quad \sigma_k^2 \equiv 0, \quad D_1 = 2\sigma_*^2 := 2\max_{x^* \in X^*} \mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x^*) - F(x^*)\|^2\right],$$
$$C = 0, \quad \rho = 1, \quad D_2 = 0.$$

□

E.2 Analysis of SGDA-AS in the Quasi-Strongly Monotone Case

Plugging the parameters from the above proposition in Theorem 2.2 and Corollary 2.3 we get the following results.

Theorem E.2. Let F be μ-quasi strongly monotone, Assumption 3.1 hold, and $0 < \gamma \leq 1/2\ell_{\mathcal{D}}$. Then, for all $k \geq 0$ the iterates produced by SGDA-AS satisfy

$$\mathbb{E}\left[\|x^k - x^{*,k}\|^2\right] \leq (1-\gamma\mu)^k\|x^0 - x^{0,*}\|^2 + \frac{2\gamma\sigma_*^2}{\mu}. \tag{47}$$

Corollary E.3 (Corollary 3.3)**.** Let the assumptions of Theorem E.2 hold. Then, for any $K \geq 0$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}
\text{if } K \leq \frac{2\ell_{\mathcal{D}}}{\mu}, &\qquad \gamma_k = \frac{1}{2\ell_{\mathcal{D}}}, \\
\text{if } K > \frac{2\ell_{\mathcal{D}}}{\mu} \text{ and } k < k_0, &\qquad \gamma_k = \frac{1}{2\ell_{\mathcal{D}}}, \\
\text{if } K > \frac{2\ell_{\mathcal{D}}}{\mu} \text{ and } k \geq k_0, &\qquad \gamma_k = \frac{2}{4\ell_{\mathcal{D}} + \mu(k-k_0)},
\end{aligned} \tag{48}$$

where $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the following inequality holds for SGDA-AS:

$$\mathbb{E}[\|x^K - x^{*,K}\|^2] \leq \frac{64\ell_{\mathcal{D}}}{\mu}\|x^0 - x^{*,0}\|^2 \exp\left(-\frac{\mu}{2\ell_{\mathcal{D}}}K\right) + \frac{72\sigma_*^2}{\mu^2 K}.$$

E.3 Analysis of SGDA-AS in the Monotone Case

In the monotone case, using Theorem 2.5, we establish the new result for SGDA-AS.

Theorem E.4. Let F be monotone ℓ-star-cocoercive and Assumptions 2.1, 2.4, 3.1 hold. Assume that $\gamma \leq 1/2\ell_{\mathcal{D}}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by SGDA-AS satisfy

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \quad &\leq \quad \frac{3\max_{u\in\mathcal{C}}\|x^0 - u\|^2}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \frac{(4\ell_{\mathcal{D}} + \ell)\|x^0 - x^{*,0}\|^2}{K} \\ &\quad + 2\gamma(2 + \gamma(4\ell_{\mathcal{D}} + \ell))\sigma_*^2 + 9\gamma \max_{x^*\in X^*}\|F(x^*)\|^2.\end{aligned}$$

Next, we apply Corollary D.4 and get the following rate of convergence to the exact solution.

Corollary E.5. Let the assumptions of Theorem E.4 hold. Then $\forall K > 0$ and

$$\gamma = \min\left\{\frac{1}{4\ell_{\mathcal{D}} + \ell}, \frac{\Omega_{0,\mathcal{C}}}{\sqrt{2K}\sigma_*}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\} \tag{49}$$

the iterates produced by SGDA-AS satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{(\ell_{\mathcal{D}} + \ell)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}(\sigma_* + G_*)}{\sqrt{K}}\right).$$

As we already mentioned before, the above result is new for SGDA-AS: the only known work on SGDA-AS (Loizou et al., 2021) focuses on the μ-quasi-strongly monotone case only with $\mu > 0$. Moreover, neglecting the dependence on problem/noise parameters, the derived convergence rate $\mathcal{O}\left(1/K + 1/\sqrt{K}\right)$ is standard for the analysis of stochastic methods for solving monotone VIPs (Juditsky et al., 2011).

E.4 Analysis of SGDA-AS in the Cocoercive Case

In the cocoercive case, using Theorem D.5, we establish the new result for SGDA-AS.

Theorem E.6. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 3.1 hold. Assume that $\gamma \leq 1/2\ell_{\mathcal{D}}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by SGDA-AS satisfy

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \quad &\leq \quad \frac{3\max_{u\in\mathcal{C}}\|x^0 - u\|^2}{2\gamma K} + \frac{(6\ell_{\mathcal{D}} + 3\ell)\|x^0 - x^{*,0}\|^2}{K} \\ &\quad + 2\gamma(3 + \gamma(6\ell_{\mathcal{D}} + 3\ell))\sigma_*^2.\end{aligned}$$

Next, we apply Corollary D.6 and get the following rate of convergence to the exact solution.

Corollary E.7. Let the assumptions of Theorem E.6 hold. Then $\forall K > 0$ and

$$\gamma = \min\left\{\frac{1}{6\ell_{\mathcal{D}} + 3\ell}, \frac{\Omega_{0,\mathcal{C}}}{\sqrt{2K}\sigma_*}\right\} \tag{50}$$

the iterates produced by SGDA-AS satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] = \mathcal{O}\left(\frac{(\ell_{\mathcal{D}}+\ell)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)}{K} + \frac{\Omega_{0,\mathcal{C}}\sigma_*}{\sqrt{K}}\right).$$

E.5 Missing Details on Arbitrary Sampling

In the main part of the paper, we discuss the Arbitrary Sampling paradigm and, in particular, using our general theoretical framework, we obtain convergence guarantees for SGDA under Expected Cocoercivity assumption (Assumption 3.1). In this section, we give the particular examples of arbitrary sampling fitting this setup. In all the examples below, we focus on a special case of stochastic reformulation from (12) and assume that for all $i \in [n]$ operator F_i is (ℓ_i, X^*)-cocoercive, i.e., for all $i \in [n]$ and $x \in \mathbb{R}^d$ we have

$$\|F_i(x) - F_i(x^*)\|^2 \leq \ell_i \langle F_i(x) - F_i(x^*), x - x^*\rangle, \tag{51}$$

where x^* is the projection of x on X^*. Note that (51) holds whenever F_i are cocoercive.

Uniform Sampling. We start with the classical uniform sampling: let $\mathbb{P}\left\{\xi = ne_i\right\} = 1/n$ for all $i \in [n]$, where $e_i \in \mathbb{R}^n$ is the i-th coordinate vector from the standard basis in $\mathbb{R}^n$. Then, $\mathbb{E}[\xi_i] = 1$ for all $i \in [n]$ and Assumption 3.1 holds with $\ell_{\mathcal{D}} = \max_{i\in[n]} \ell_i$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x) - F_\xi(x^*)\|^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\|F_i(x) - F_i(x^*)\|^2 \\
&\overset{(51)}{\leq} \frac{1}{n}\sum\left(\ell_i\langle F_i(x) - F_i(x^*), x - x^*\rangle\right) \\
&\leq \max_{i\in[n]}\ell_i\langle F(x) - F(x^*), x - x^*\rangle
\end{aligned}$$

In this case, Corollaries E.3 and E.5 imply the following rate for SGDA in μ-quasi strongly monotone, monotone and cocoercive cases respectively:

$$\begin{aligned}
\mathbb{E}[\|x^K - x^{*,K}\|^2] &\leq \frac{64\max_{i\in[n]}\ell_i}{\mu}\|x^0 - x^{*,0}\|^2\exp\left(-\frac{\mu}{2\max_{i\in[n]}\ell_i}K\right) + \frac{72\sigma_{*,\mathrm{US}}^2}{\mu^2 K}, \\
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &= \mathcal{O}\left(\frac{(\max_{i\in[n]}\ell_i + \ell)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}(\sigma_{*,\mathrm{US}} + G_*)}{\sqrt{K}}\right), \\
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &= \mathcal{O}\left(\frac{(\max_{i\in[n]}\ell_i + \ell)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)}{K} + \frac{\Omega_{0,\mathcal{C}}\sigma_{*,\mathrm{US}}}{\sqrt{K}}\right),
\end{aligned}$$

where $\sigma_{*,\mathrm{US}}^2 := \max_{x^*\in X^*}\frac{1}{n}\sum_{i=1}^{n}\|F_i(x^*) - F(x^*)\|^2$.

Importance Sampling. Next, we consider a non-uniform sampling strategy – importance sampling: let $\mathbb{P}\left\{\xi = e_i n\overline{\ell}/\ell_i\right\} = \ell_i/n\overline{\ell}$ for all $i \in [n]$, where $\overline{\ell} = \frac{1}{n}\sum_{i=1}^{n}\ell_i$. Then, $\mathbb{E}[\xi_i] = 1$ for all $i \in [n]$ and Assumption 3.1 holds with $\ell_{\mathcal{D}} = \overline{\ell}$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\|F_\xi(x) - F_\xi(x^*)\|^2\right] &= \sum_{i=1}^{n}\frac{\ell_i}{n\overline{\ell}}\left\|\frac{\overline{\ell}}{\ell_i}\left(F_i(x) - F_i(x^*)\right)\right\|^2 \\
&= \sum_{i=1}^{n}\frac{\overline{\ell}}{n\ell_i}\|F_i(x) - F_i(x^*)\|^2 \\
&\overset{(51)}{\leq} \frac{\overline{\ell}}{n}\sum\langle F_i(x) - F_i(x^*), x - x^*\rangle \\
&\leq \overline{\ell}\langle F(x) - F(x^*), x - x^*\rangle
\end{aligned}$$

In this case, Corollaries E.3 and E.5 imply the following rate for SGDA in μ-quasi strongly monotone, monotone and cocoercive cases respectively:

$$\begin{aligned}
\mathbb{E}[\|x^K - x^{*,K}\|^2] &\leq \frac{64\bar{\ell}}{\mu}\|x^0 - x^{*,0}\|^2 \exp\left(-\frac{\mu}{2\bar{\ell}}K\right) + \frac{72\sigma_{*,\text{IS}}^2}{\mu^2 K},\\
\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] &= \mathcal{O}\left(\frac{(\bar{\ell}+\ell)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}(\sigma_{*,\text{IS}} + G_*)}{\sqrt{K}}\right),\\
\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] &= \mathcal{O}\left(\frac{(\bar{\ell}+\ell)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2)}{K} + \frac{\Omega_{0,\mathcal{C}}\sigma_{*,\text{IS}}}{\sqrt{K}}\right)
\end{aligned}$$

where $\sigma_{*,\text{IS}}^2 := \max_{x^*\in X^*}\frac{1}{n}\sum_{i=1}^n \frac{\ell_i}{\bar{\ell}}\left\|\frac{\bar{\ell}}{\ell_i}F_i(x^*) - F(x^*)\right\|^2$. We emphasize that $\bar{\ell} \leq \max_{i\in[n]}\ell_i$ and, in fact, $\bar{\ell}$ might be much smaller than $\max_{i\in[n]}\ell_i$. Therefore, compared to SGDA with uniform sampling, SGDA with importance sampling has better exponentially decaying term in the quasi-strongly monotone case and converges faster to the neighborhood, if executed with constant stepsize. Moreover, $\sigma_{*,\text{IS}}^2 \leq \sigma_{*,\text{US}}^2$, when $\max_{x^*\in X^*}\|F_i(x^*)\| \sim \ell_i$. In this case, SGDA with importance sampling has better $\mathcal{O}(1/K)$ term than SGDA with uniform sampling as well.

Minibatch Sampling With Replacement. Let $\xi = \frac{1}{b}\sum_{i=1}^b \xi^i$, where ξ^i are i.i.d. samples from some distribution $\mathcal{D}$ satisfying (12) and Assumption 3.1. Then, the distribution of ξ satisfies (12) and Assumption 3.1 as well with the same constant $\ell_{\mathcal{D}}$. Therefore, minibatched versions of uniform sampling and importance sampling fit the framework as well with $\ell_{\mathcal{D}} = \max_{i\in[n]}\ell_i, \sigma_*^2 = \frac{\sigma_{*,\text{US}}^2}{b}$ and $\ell_{\mathcal{D}} = \bar{\ell}, \sigma_*^2 = \frac{\sigma_{*,\text{IS}}^2}{b}$.

Minibatch Sampling Without Replacement. For given batchsize $b \in [n]$ we consider the following sampling strategy: for each subset $S \subseteq [n]$ such that $|S| = b$ we have $\mathbb{P}\left\{\xi = \frac{n}{b}\sum_{i\in S} e_i\right\} = \frac{b!(n-b)!}{n!}$, i.e., S is chosen uniformly at random from all b-element subsets of $[n]$. In the special case, when $R(x) \equiv 0$, Loizou et al. (2021) show that this sampling strategy satisfies (12) and Assumption 3.1 with

$$\ell_{\mathcal{D}} = \frac{n(b-1)}{b(n-1)}\ell + \frac{n-b}{b(n-1)}\max_{i\in[n]}\ell_i, \quad \sigma_*^2 = \frac{n-b}{b(n-1)}\sigma_{*,\text{US}}^2. \tag{52}$$

Clearly, both parameters are smaller than corresponding parameters for minibatched version of uniform sampling with replacement, which indicates the theoretical benefits of sampling without replacement. Plugging the parameters from (52) in Corollaries E.3 and E.5, we get the rate of convergence for this sampling strategy. Moreover, in the quasi-strongly monotone case, to guarantee $\mathbb{E}[\|x^K - x^{*,K}\|^2] \leq \varepsilon$ for some $\varepsilon > 0$, the method requires

$$\begin{aligned}
Kb &= \mathcal{O}\left(\max\left\{\left(b\frac{\ell}{\mu} + \frac{(n-b)}{n}\frac{\max_{i\in[n]}\ell_i}{\mu}\right)\log\frac{\ell_{\mathcal{D}}\|x^0 - x^{*,0}\|^2}{\mu\varepsilon}, \frac{(n-b)\sigma_{*,\text{US}}^2}{n\mu^2\varepsilon}\right\}\right)\\
&= \widetilde{\mathcal{O}}\left(\max\left\{\frac{b\left(\ell - \frac{1}{n}\max_{i\in[n]}\ell_i\right) + \max_{i\in[n]}\ell_i}{\mu}, \frac{(n-b)\sigma_{*,\text{US}}^2}{n\mu^2\varepsilon}\right\}\right) \quad \text{oracle calls,}
\end{aligned} \tag{53}$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides numerical and logarithmic factors. One can notice that the first term in the maximum linearly increases in b (since ℓ cannot be smaller than $\frac{1}{n}\max_{i\in[n]}\ell_i$), while the second term linearly decreases in b. The first term in the maximum is lower bounded by $\frac{(n-b)}{n}\frac{\max_{i\in[n]}\ell_i}{\mu}$. Therefore, if $\max_{i\in[n]}\ell_i \geq \frac{\sigma_{*,\text{US}}^2}{\mu\varepsilon}$, the the first term in the maximum is always larger than the second one, meaning that the optimal batchsize, i.e., the batchsize that minimizes oracle complexity (53) neglecting the logarithmic terms, equals $b_* = 1$. Next, if $\max_{i\in[n]}\ell_i < \frac{\sigma_{*,\text{US}}^2}{\mu\varepsilon}$, then there exists a positive value of b such that the first term in the maximum equals the second term. This value equals

$$\frac{n\left(\sigma_{*,\text{US}}^2 - \mu\varepsilon\max_{i\in[n]}\ell_i\right)}{\sigma_*^2 + \mu\varepsilon\left(n\ell - \max_{i\in[n]}\ell_i\right)}.$$

One can easily verify that it is always smaller than n, but it can be non integer and it can be smaller than 1 as well. Therefore, the optimal batchsize is

$$b_* = \begin{cases} 1, & \text{if } \max_{i\in[n]}\ell_i \geq \frac{\sigma_{*,\text{US}}^2}{\mu\varepsilon},\\ \max\left\{1, \left\lfloor\frac{n\left(\sigma_{*,\text{US}}^2 - \mu\varepsilon\max_{i\in[n]}\ell_i\right)}{\sigma_*^2 + \mu\varepsilon\left(n\ell - \max_{i\in[n]}\ell_i\right)}\right\rfloor\right\}, & \text{otherwise.}\end{cases}$$

We notice that Loizou et al. (2021) derive the following formula for the optimal batchsize (ignoring numerical constants):

$$\widetilde{b}_* = \begin{cases} 1, & \text{if } \max_{i\in[n]} \ell_i - \ell \geq \frac{\sigma^2_{*,\mathrm{US}}}{\mu\varepsilon}, \\ \max\left\{1, \left\lfloor \frac{n\left(\sigma^2_{*,\mathrm{US}} - \mu\varepsilon(\max_{i\in[n]}\ell_i - \ell)\right)}{\sigma^2_* + \mu\varepsilon\left(n\ell - \max_{i\in[n]}\ell_i\right)} \right\rfloor\right\}, & \text{otherwise.} \end{cases}$$

However, in terms of $\widetilde{\mathcal{O}}(\cdot)$ both formulas give the same complexity result.

F SGDA WITH VARIANCE REDUCTION: MISSING PROOFS AND DETAILS

In this section, we provide missing proofs and details for Section 4.

F.1 L-SVRGDA

Algorithm 2 L-SVRGDA: Loopless Stochastic Variance Reduced Gradient Descent-Ascent

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, probability $p \in (0, 1]$, stepsize $\gamma > 0$, number of steps K
2: Set $w^0 = x^0$ and compute $F(w^0)$
3: **for** $k = 0$ **to** $K - 1$ **do**
4: Draw a fresh sample j_k from the uniform distribution on $[n]$ and compute $g^k = F_{j_k}(x^k) - F_{j_k}(w^k) + F(w^k)$
5: $w^{k+1} = \begin{cases} x^k, & \text{with probability } p, \\ w^k, & \text{with probability } 1-p, \end{cases}$
6: $x^{k+1} = \text{prox}_{\gamma R}(x^k - \gamma g^k)$
7: **end for**

F.1.1 Proof of Proposition 4.3

Lemma F.1. Let Assumption 4.1 hold. Then for all $k \geq 0$ L-SVRGDA satisfies

$$\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] \leq 2\widehat{\ell}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + 2\sigma_k^2, \tag{54}$$

where $\sigma_k^2 := \frac{1}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^{*,k})\|^2$.

Proof. Since $g^k = F_{j_k}(x^k) - F_{j_k}(w^k) + F(w^k)$, we have

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &= \mathbb{E}_k\left[\|F_{j_k}(x^k) - F_{j_k}(w^k) + F(w^k) - F(x^{*,k})\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(w^k) + F(w^k) - F(x^{*,k})\|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^{*,k})\|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^{*,k}) - (F(w^k) - F(x^{*,k}))\|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^{*,k})\|^2 + \frac{2}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^{*,k})\|^2 \\
&\overset{(14)}{\leq} 2\widehat{\ell}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + 2\sigma_k^2.
\end{aligned}$$

□

Lemma F.2. Let Assumptions 4.1 and 4.2 hold. Then for all $k \geq 0$ L-SVRGDA satisfies

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq p\widehat{\ell}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + (1-p)\sigma_k^2, \tag{55}$$

where $\sigma_k^2 := \frac{1}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^{*,k})\|^2$.

Proof. Using the definitions of σ_{k+1}^2 and w^{k+1} (see (13)), we derive

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\|F_i(w^{k+1}) - F_i(x^{*,k+1})\|^2\right]$$

$$\begin{aligned}
&\overset{\text{As. 4.2}}{=} \quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\|F_i(w^{k+1}) - F_i(x^{*,k})\|^2\right] \\
&= \quad \frac{p}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^{*,k})\|^2 + \frac{1-p}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^{*,k})\|^2 \\
&\overset{(14)}{\leq} \quad p\widehat{\ell}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + (1-p)\sigma_k^2.
\end{aligned}$$

□

The above two lemmas imply that Assumption 2.1 is satisfied with certain parameters.

Proposition F.3 (Proposition 4.3)**.** Let Assumptions 4.1 and 4.2 hold. Then, L-SVRGDA satisfies Assumption 2.1 with

$$A = \widehat{\ell}, \quad B = 2, \quad \sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|F_i(w^k) - F_i(x^*)\|^2, \quad C = \frac{p\widehat{\ell}}{2}, \quad \rho = p, \quad D_1 = D_2 = 0.$$

F.1.2 Analysis of L-SVRGDA in the Quasi-Strongly Monotone Case

Plugging the parameters from the above proposition in Theorem 2.2 and Corollary 2.3 with $M = \frac{4}{p}$ we get the following results.

Theorem F.4. Let F be μ-quasi strongly monotone, Assumptions 4.1, 4.2 hold, and $0 < \gamma \leq 1/6\widehat{\ell}$. Then for all $k \geq 0$ the iterates produced by L-SVRGDA satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, p/2\right\}\right)^k V_0, \tag{56}$$

where $V_0 = \|x^0 - x^*\|^2 + 4\gamma^2\sigma_0^2/p$.

Corollary F.5. Let the assumptions of Theorem F.4 hold. Then, for $p = n$, $\gamma = 1/6\widehat{\ell}$ and any $K \geq 0$ we have

$$\mathbb{E}[\|x^k - x^*\|^2] \leq V_0 \exp\left(-\min\left\{\frac{\mu}{6\widehat{\ell}}, \frac{1}{2n}\right\} K\right).$$

F.1.3 Analysis of L-SVRGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of L-SVRGDA in the monotone case.

Theorem F.6. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2 hold. Assume that $\gamma \leq 1/6\widehat{\ell}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates of L-SVRGDA satisfy

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \quad &\leq \quad \frac{3\max_{u\in\mathcal{C}}\|x^0 - u\|^2}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \frac{\left(12\widehat{\ell} + \ell\right)\|x^0 - x^{*,0}\|^2}{K} \\
&\quad + \left(4 + \left(12\widehat{\ell} + \ell\right)\gamma\right)\frac{2\gamma\sigma_0^2}{pK} + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2.
\end{aligned}$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary F.7. Let the assumptions of Theorem F.6 hold and $p = 1/n$. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{12\widehat{\ell} + \ell}, \frac{1}{\sqrt{2n\ell\widehat{\ell}}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}. \tag{57}$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] = \mathcal{O}\left(\frac{(\widehat{\ell}+\ell)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)+\sqrt{n\widehat{\ell}\ell}\Omega_{0,\mathcal{C}}^2+\ell\Omega_{\mathcal{C}}^2}{K}+\frac{\Omega_{0,\mathcal{C}}G_*}{\sqrt{K}}\right).$$

Proof. First of all, (14), (4), and Cauchy-Schwarz inequality imply

$$\begin{aligned}
\sigma_0^2 &= \frac{1}{n}\sum_{i=1}^{n}\|F_i(x^0)-F_i(x^*)\|^2 \\
&\overset{(14)}{\le} \widehat{\ell}\langle F(x^0)-F(x^*), x^0-x^*\rangle \\
&\le \widehat{\ell}\|F(x^0)-F(x^*)\|\cdot\|x^0-x^*\| \\
&\le \widehat{\ell}\ell\|x^0-x^*\|^2 \le \widehat{\ell}\ell\max_{u\in\mathcal{C}}\|x^0-u\|^2 \le \widehat{\ell}\ell\Omega_{0,\mathcal{C}}^2.
\end{aligned}$$

Next, applying Corollary D.4 with $\widehat{\sigma}_0 := \sqrt{\widehat{\ell}\ell}\Omega_{0,\mathcal{C}}$, we get the result. □

F.1.4 Analysis of L-SVRGDA in the Cocoercive Case

Next, using Theorem 2.6, we establish the convergence of L-SVRGDA in the cocoercive case.

Theorem F.8. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2 hold. Assume that $\gamma \le 1/6\widehat{\ell}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \ge 0$ the iterates of L-SVRGDA satisfy

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] &\le \frac{3\max_{u\in\mathcal{C}}\|x^0-u\|^2}{2\gamma K}+\frac{\left(18\widehat{\ell}+3\ell\right)\|x^0-x^{*,0}\|^2}{K} \\
&\quad + \left(6+\left(18\widehat{\ell}+3\ell\right)\gamma\right)\frac{2\gamma\sigma_0^2}{pK}.
\end{aligned}$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary F.9. Let the assumptions of Theorem F.8 hold and $p = 1/n$. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{18\widehat{\ell}+3\ell}, \frac{1}{\sqrt{2n\widehat{\ell}\ell}}\right\}. \tag{58}$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right] = \mathcal{O}\left(\frac{(\widehat{\ell}+\ell)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)+\sqrt{n\widehat{\ell}\ell}\Omega_{0,\mathcal{C}}^2}{K}\right).$$

F.2 SAGA-SGDA

In this section, we show that SAGA-SGDA (Palaniappan and Bach, 2016) fits our theoretical framework and derive new results for this method under averaged star-cocoercivity.

F.2.1 SAGA-SGDA Fits Assumption 2.1

Lemma F.10. Let Assumption 4.1 hold. Then for all $k \ge 0$ SAGA-SGDA satisfies

$$\mathbb{E}_k\left[\|g^k-F(x^{*,k})\|^2\right] \le 2\widehat{\ell}\langle F(x^k)-F(x^{*,k}), x^k-x^{*,k}\rangle + 2\sigma_k^2, \tag{59}$$

where $\sigma_k^2 := \frac{1}{n}\sum_{i=1}^{n}\|F_i(w_i^k)-F_i(x^{*,k})\|^2$.

Algorithm 3 SAGA-SGDA (Palaniappan and Bach, 2016)

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, number of steps K
2: Set $w_i^0 = x^0$ and compute $F_i(w_i^0)$ for all $i \in [n]$
3: **for** $k = 0$ **to** $K-1$ **do**
4: Draw a fresh sample j_k from the uniform distribution on $[n]$ and compute $g^k = F_{j_k}(x^k) - F_{j_k}(w_{j_k}^k) + \frac{1}{n}\sum_{i=1}^n F_i(w_i^k)$
5: Set $w_{j_k}^{k+1} = x^k$ and $w_i^{k+1} = w_i^k$ for $i \neq j_k$
6: $x^{k+1} = \mathrm{prox}_{\gamma R}(x^k - \gamma g^k)$
7: **end for**

Proof. For brevity, we introduce a new notation: $S^k = \frac{1}{n}\sum_{i=1}^n F_i(w_i^k)$. Since $g^k = F_{j_k}(x^k) - F_{j_k}(w_{j_k}^k) + S^k$, we have

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &= \mathbb{E}_k\left[\|F_{j_k}(x^k) - F_{j_k}(w_{j_k}^k) + S^k - F(x^{*,k})\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(w_i^k) + S^k - F(x^{*,k})\|^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^{*,k})\|^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^n \|F_i(w_i^k) - F_i(x^{*,k}) - (S^k - F(x^{*,k}))\|^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^{*,k})\|^2 + \frac{2}{n}\sum_{i=1}^n \|F_i(w_i^k) - F_i(x^{*,k})\|^2 \\
&\overset{(14)}{\le} 2\widehat{\ell}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + 2\sigma_k^2.
\end{aligned}$$

□

Lemma F.11. Let Assumptions 4.1 and 4.2 hold. Then for all $k \ge 0$ SAGA-SGDA satisfies

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \le \frac{\widehat{\ell}}{n}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + (1 - 1/n)\sigma_k^2, \tag{60}$$

where $\sigma_k^2 := \frac{1}{n}\sum_{i=1}^n \|F_i(w_i^k) - F_i(x^{*,k})\|^2$.

Proof. Using the definitions of σ_{k+1}^2 and w_i^{k+1}, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\|F_i(w_i^{k+1}) - F_i(x^{*,k+1})\|^2\right] \\
&\overset{\text{As. } 4.2}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\|F_i(w_i^{k+1}) - F_i(x^{*,k})\|^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \|F_i(x^k) - F_i(x^{*,k})\|^2 + \frac{1 - 1/n}{n}\sum_{i=1}^n \|F_i(w_i^k) - F_i(x^{*,k})\|^2 \\
&\overset{(14)}{\le} \frac{\widehat{\ell}}{n}\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle + (1 - 1/n)\sigma_k^2.
\end{aligned}$$

□

The above two lemmas imply that Assumption 2.1 is satisfied with certain parameters.

Proposition F.12. Let Assumptions 4.1 and 4.2 hold. Then, SAGA-SGDA satisfies Assumption 2.1 with

$$A = \widehat{\ell}, \quad B = 2, \quad \sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|F_i(w_i^k) - F_i(x^*)\|^2, \quad C = \frac{\widehat{\ell}}{2n}, \quad \rho = \frac{1}{n}, \quad D_1 = D_2 = 0.$$

F.2.2 Analysis of SAGA-SGDA in the Quasi-Strongly Monotone Case

Applying Theorem 2.2 and Corollary 2.3 with $M = 4n$, we get the following results.

Theorem F.13. Let F be μ-quasi strongly monotone, Assumptions 4.1, 4.2 hold, and $0 < \gamma \leq 1/6\widehat{\ell}$. Then for all $k \geq 0$ the iterates produced by SAGA-SGDA satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, 1/2n\right\}\right)^k V_0, \tag{61}$$

where $V_0 = \|x^0 - x^*\|^2 + 4n\gamma^2\sigma_0^2$.

Corollary F.14. Let the assumptions of Theorem F.13 hold. Then, for $\gamma = 1/6\widehat{\ell}$ and any $K \geq 0$ we have

$$\mathbb{E}[\|x^K - x^*\|^2] \leq V_0 \exp\left(-\min\left\{\frac{\mu}{6\widehat{\ell}}, \frac{1}{2n}\right\}K\right).$$

F.2.3 Analysis of SAGA-SGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of SAGA-SGDA in the monotone case.

Theorem F.15. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2 hold. Assume that $\gamma \leq 1/6\widehat{\ell}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by SAGA-SGDA satisfy

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \quad &\leq \quad \frac{3\max_{u\in\mathcal{C}}\|x^0 - u\|^2}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \frac{\left(12\widehat{\ell} + \ell\right)\|x^0 - x^{*,0}\|^2}{K} \\ &\quad + \left(4 + \left(12\widehat{\ell} + \ell\right)\gamma\right)\frac{2\gamma\sigma_0^2}{pK} + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2.\end{aligned}$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary F.16. Let the assumptions of Theorem F.15 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{12\widehat{\ell} + \ell}, \frac{1}{\sqrt{2n\ell\widehat{\ell}}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}, \tag{62}$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] = \mathcal{O}\left(\frac{(\widehat{\ell} + \ell)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \sqrt{n\ell\widehat{\ell}}\Omega_{0,\mathcal{C}}^2 + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}G_*}{\sqrt{K}}\right).$$

Proof. Since σ_0 for SAGA-SGDA and L-SVRGDA are the same, the proof of this corollary is identical to the one for Corollary F.7. □

F.2.4 Analysis of SAGA-SGDA in the Cocoercive Case

Next, using Theorem 2.6, we establish the convergence of SAGA-SGDA in the cocoercive case.

Theorem F.17. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2 hold. Assume that $\gamma \leq 1/6\widehat{\ell}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$

from (9) and for all $K \geq 0$ the iterates produced by SAGA-SGDA satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \leq \frac{3\max_{u\in\mathcal{C}}\|x^0-u\|^2}{2\gamma K} + \frac{\left(18\widehat{\ell}+3\ell\right)\|x^0-x^{*,0}\|^2}{K} + \left(6+\left(18\widehat{\ell}+3\ell\right)\gamma\right)\frac{2\gamma\sigma_0^2}{pK}.$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary F.18. Let the assumptions of Theorem F.17 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{18\widehat{\ell}+3\ell}, \frac{1}{\sqrt{2n\ell\widehat{\ell}}}\right\}, \tag{63}$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{(\widehat{\ell}+\ell)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)+\sqrt{n\widehat{\ell}\ell}\,\Omega_{0,\mathcal{C}}^2}{K}\right).$$

F.3 Discussion of the Results in the Monotone and Cocoercive Cases

Among the papers mentioned in the related work on variance-reduced methods (see Section A), only Alacaoglu and Malitsky (2021); Carmon et al. (2019); Alacaoglu et al. (2021); Tominin et al. (2021); Luo et al. (2021) consider monotone (convex-concave) and Lipschitz (smooth) VIPs (min-max problems) without assuming strong monotonicity (strong-convexity-strong-concavity) of the problem. In this case, Alacaoglu and Malitsky (2021) derive $\mathcal{O}\left(n+\frac{\sqrt{n}L}{K}\right)$ convergence rate (neglecting the dependence on the quantities like $\Omega_{0,\mathcal{C}}^2 = \max_{u\in\mathcal{C}}\|x^0-u\|^2$), which is optimal for the considered setting (Han et al., 2021). Under additional assumptions a similar rate is derived in Carmon et al. (2019). Tominin et al. (2021); Luo et al. (2021) also achieve this rate but using Catalyst. Finally, Alacaoglu et al. (2021) derive $\mathcal{O}\left(n+\frac{nL}{K}\right)$, which is worse than the one from Alacaoglu and Malitsky (2021). Our results for monotone and star-cocoercive regularized VIPs give $\mathcal{O}\left(\frac{\sqrt{n\ell\widehat{\ell}}+\widehat{\ell}}{K}+\frac{G_*}{\sqrt{K}}\right)$ rate, which is typically worse than $\mathcal{O}\left(n+\frac{\sqrt{n}L}{K}\right)$ rate from Alacaoglu and Malitsky (2021) due to the relation between cocoercivity constants and Lipschitz constants (even when $R(x)\equiv 0$, i.e., $G_*=0$). However, in general, it is possible that star-cocoercivity holds, while Lipschitzness does not (Loizou et al., 2021). As for cocoercive case, we obtain $\mathcal{O}\left(\frac{\sqrt{n\ell\widehat{\ell}}+\widehat{\ell}}{K}\right)$, which matches the rate from Alacaoglu and Malitsky (2021) up to the difference between cocoercivity and Lipschitz constants. Moreover, we emphasize here that Alacaoglu and Malitsky (2021) and other works do not consider SGDA as the basis for their methods. To the best of our knowledge, our results are the first ones for variance-reduced SGDA-type methods derived in the monotone case without assuming (quasi-)strong monotonicity.

G DISTRIBUTED SGDA WITH COMPRESSION: MISSING PROOFS AND DETAILS

In this section, we provide missing proofs and details for Section 5.

G.1 QSGDA

In this section (and in the one about DIANA-SGDA), we assume that each F_i has an expectation form: $F_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[F_{\xi_i}(x)]$.

Algorithm 4 QSGDA: Quantized Stochastic Gradient Descent-Ascent

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, number of steps K
2: **for** $k = 0$ **to** $K-1$ **do**
3: Broadcast x^k to all workers
4: **for** $i = 1, \ldots, n$ in parallel **do**
5: Compute g_i^k and send $\mathcal{Q}(g_i^k)$ to the server
6: **end for**
7: $g^k = \frac{1}{n}\sum_{i=1}^n \mathcal{Q}(g_i^k)$
8: $x^{k+1} = \text{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
9: **end for**

G.1.1 Proof of Proposition 5.3

Proposition G.1 (Proposition 5.3)**.** Let F be ℓ-star-cocoercive and Assumptions 4.1, 5.2 hold. Then, QSGDA with quantization (15) satisfies Assumption 2.1 with

$$A = \left(\frac{3\ell}{2} + \frac{9\omega\widehat{\ell}}{2n}\right), \quad D_1 = \frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n}, \quad \sigma_k^2 = 0, \quad B = 0,$$
$$C = 0, \quad \rho = 1, \quad D_2 = 0,$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ and $\zeta_*^2 = \frac{1}{n}\max_{x*\in X^*}\left[\sum_{i=1}^n \|F_i(x^*)\|^2\right]$.

Proof. Since $g^k = \frac{1}{n}\sum\limits_{i=1}^n \mathcal{Q}\left(g_i^k\right)$, $\mathcal{Q}\left(g_1^k\right), \ldots, \mathcal{Q}\left(g_n^k\right)$ are independent for fixed $g_1^k, \ldots, g_n^k$, and $g_1^k, \ldots, g_n^k$ are independent for fixed x^k, we have

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &= \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}\left(g_i^k\right) - F(x^{*,k})\right\|^2\right] \\
&= \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n \left[\mathcal{Q}\left(g_i^k\right) - g_i^k + g_i^k - F_i(x^k)\right] + F(x^k) - F(x^{*,k})\right\|^2\right] \\
&\leq 3\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n [\mathcal{Q}\left(g_i^k\right) - g_i^k]\right\|^2\right] + 3\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n [g_i^k - F_i(x^k)]\right\|^2\right] \\
&\quad +3\left\|F(x^k) - F(x^{*,k})\right\|^2 \\
&= \frac{3}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|\mathcal{Q}\left(g_i^k\right) - g_i^k\right\|^2\right] + \frac{3}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|g_i^k - F_i(x^k)\right\|^2\right] \\
&\quad +3\left\|F(x^k) - F(x^{*,k})\right\|^2.
\end{aligned}$$

Next, we use Assumption 5.2, $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$, and the definition of quantization (15) and get

$$\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] \leq \frac{3\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|g_i^k\right\|^2\right] + \frac{3\sigma^2}{n} + 3\left\|F(x^k) - F(x^{*,k})\right\|^2$$

$$
\begin{aligned}
&\leq \frac{3\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|g_i^k - F_i(x^k) + F_i(x^k) - F_i(x^{*,k}) + F_i(x^{*,k})\right\|^2\right] \\
&\quad + \frac{3\sigma^2}{n} + 3\left\|F(x^k) - F(x^{*,k})\right\|^2 \\
&\leq \frac{9\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|g_i^k - F_i(x^k)\right\|^2\right] + \frac{9\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|F_i(x^k) - F_i(x^{*,k})\right\|^2\right] \\
&\quad + \frac{9\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|F_i(x^{*,k})\right\|^2\right] + \frac{3\sigma^2}{n} + 3\left\|F(x^k) - F(x^{*,k})\right\|^2 \\
&\overset{(16)}{\leq} \frac{9\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|F_i(x^k) - F_i(x^{*,k})\right\|^2\right] + 3\left\|F(x^k) - F(x^{*,k})\right\|^2 \\
&\quad + \frac{9\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|F_i(x^{*,k})\right\|^2\right] + \frac{3(1+3\omega)\sigma^2}{n}.
\end{aligned}
$$

Star-cocoercivity of F and Assumption 4.1 give

$$
\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &\leq \left(3\ell + \frac{9\omega}{n}\widehat{\ell}\right)\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle \\
&\quad + \frac{9\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|F_i(x^{*,k})\right\|^2\right] + \frac{3(1+3\omega)\sigma^2}{n} \\
&\leq \left(3\ell + \frac{9\omega}{n}\widehat{\ell}\right)\langle F(x^k) - F(x^{*,k}), x^k - x^{*,k}\rangle \\
&\quad + \frac{9\omega}{n^2}\max_{x*\in X^*}\left[\sum_{i=1}^n \|F_i(x^*)\|^2\right] + \frac{3(1+3\omega)\sigma^2}{n}.
\end{aligned}
$$

□

G.1.2 Analysis of QSGDA in the Quasi-Strongly Monotone Case

Applying Theorem 2.2 and Corollary 2.3, we get the following results.

Theorem G.2. Let F be μ-quasi strongly monotone, ℓ-star-cocoercive, Assumptions 4.1, 5.2 hold, and

$$
0 < \gamma \leq \frac{1}{3\ell + \frac{9\omega\widehat{\ell}}{n}}.
$$

Then, for all $k \geq 0$ the iterates produced by QSGDA satisfy

$$
\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k\|x^0 - x^*\|^2 + \gamma\frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n\mu}.
$$

Corollary G.3. Let the assumptions of Theorem G.2 hold. Then, for any $K \geq 0$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$
\begin{aligned}
&\text{if } K \leq \frac{1}{\mu}\cdot\left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right), && \gamma_k = \left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right)^{-1}, \\
&\text{if } K > \frac{1}{\mu}\cdot\left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right) \text{ and } k < k_0, && \gamma_k = \left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right)^{-1}, \\
&\text{if } K > \frac{1}{\mu}\cdot\left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right) \text{ and } k \geq k_0, && \gamma_k = \frac{2}{(6\ell + 18\omega\widehat{\ell}/n + \mu(k-k_0))},
\end{aligned}
$$

where $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the following inequality holds:

$$\mathbb{E}[\|x^K - x^{*,K}\|^2] \leq \frac{32(3\ell + 9\omega\widehat{\ell}/n)}{\mu}\|x^0 - x^{*,0}\|^2 \exp\left(-\frac{\mu}{(3\ell + 9\omega\widehat{\ell}/n)}K\right) + \frac{36}{\mu^2 K} \cdot \frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n}.$$

G.1.3 Analysis of QSGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of QSGDA in the monotone case.

Theorem G.4. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4, 4.1, 5.2 hold. Assume that $\gamma \leq \left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right)^{-1}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by QSGDA satisfy

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \leq\; & \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \left(7\ell + \frac{18\omega\widehat{\ell}}{n}\right)\cdot\frac{\|x^0 - x^{*,0}\|^2}{K} \\ & + \gamma\left(2 + \gamma\left(7\ell + \frac{18\omega\widehat{\ell}}{n}\right)\right)\cdot\frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n} \\ & + 9\gamma \max_{x^*\in X^*}\left[\|F(x^*)\|^2\right]\end{aligned}$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary G.5. Let the assumptions of Theorem G.4 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{7\ell + \frac{18\omega\widehat{\ell}}{n}}, \frac{\Omega_{0,\mathcal{C}}\sqrt{n}}{\sqrt{3K(1+3\omega)\sigma^2 + 9K\omega\zeta_*^2}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}.$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] = \mathcal{O}\left(\frac{(\ell + \omega\widehat{\ell}/n)\left(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2\right) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}(\sigma\sqrt{1+\omega} + G_*\sqrt{n} + \zeta_*\sqrt{\omega)}}{\sqrt{nK}}\right).$$

G.1.4 Analysis of QSGDA in the Cocoercive Case

Next, using Theorem 2.6, we establish the convergence of QSGDA in the cocoercive case.

Theorem G.6. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 4.1, 5.2 hold. Assume that $\gamma \leq \left(3\ell + \frac{9\omega\widehat{\ell}}{n}\right)^{-1}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by QSGDA satisfy

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \leq\; & \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \left(10\ell + \frac{27\omega\widehat{\ell}}{n}\right)\cdot\frac{\|x^0 - x^{*,0}\|^2}{K} \\ & + \gamma\left(3 + \gamma\left(10\ell + \frac{27\omega\widehat{\ell}}{n}\right)\right)\cdot\frac{3(1+3\omega)\sigma^2 + 9\omega\zeta_*^2}{n}.\end{aligned}$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary G.7. Let the assumptions of Theorem G.6 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{10\ell + \frac{27\omega\widehat{\ell}}{n}}, \frac{\Omega_{0,\mathcal{C}}\sqrt{n}}{\sqrt{3K(1+3\omega)\sigma^2 + 9K\omega\zeta_*^2}}\right\}.$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{\left(\ell + {\omega\widehat{\ell}}/{n}\right)\left(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2\right)}{K} + \frac{\Omega_{0,\mathcal{C}}(\sigma\sqrt{1+\omega} + \zeta_*\sqrt{\omega})}{\sqrt{nK}}\right).$$

G.2 DIANA-SGDA

Algorithm 5 DIANA-SGDA: DIANA Stochastic Gradient Descent-Ascent Mishchenko et al. (2019); Horváth et al. (2019)

1: **Input:** starting points $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$, $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$, stepsizes $\gamma, \alpha > 0$, number of steps K
2: **for** $k = 0$ **to** $K-1$ **do**
3: Broadcast x^k to all workers
4: **for** $i = 1, \ldots, n$ in parallel **do**
5: Compute g_i^k and $\Delta_i^k = g_i^k - h_i^k$
6: Send $\mathcal{Q}(\Delta_i^k)$ to the server
7: $h_i^{k+1} = h_i^k + \alpha\mathcal{Q}(\Delta_i^k)$
8: **end for**
9: $g^k = h^k + \frac{1}{n}\sum\limits_{i=1}^n \mathcal{Q}(\Delta_i^k) = \frac{1}{n}\sum\limits_{i=1}^n (h_i^k + \mathcal{Q}(\Delta_i^k))$
10: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
11: $h^{k+1} = h^k + \alpha\frac{1}{n}\sum\limits_{i=1}^n \mathcal{Q}(\Delta_i^k) = \frac{1}{n}\sum\limits_{i=1}^n h_i^k$
12: **end for**

G.2.1 Proof of Proposition 5.4

The following result follows from Lemmas 1 and 2 from Horváth et al. (2019). It holds in our settings as well, since it does not rely on the exact form of $F_i(x^k)$.

Lemma G.8 (Lemmas 1 and 2 from Horváth et al. (2019))**.** Let Assumptions 4.2, 5.2 hold. Suppose that $\alpha \leq {1}/{(1+\omega)}$. Then, for all $k \geq 0$ DIANA-SGDA satisfies

$$\begin{aligned}
\mathbb{E}_k\left[g^k\right] &= F(x^k),\\
\mathbb{E}_k\left[\|g^k - F(x^*)\|^2\right] &\leq \left(1 + \frac{2\omega}{n}\right)\frac{1}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^*)\|^2 + \frac{2\omega\sigma_k^2}{n} + \frac{(1+\omega)\sigma^2}{n},\\
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &\leq (1-\alpha)\sigma_k^2 + \frac{\alpha}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^*)\|^2 + \alpha\sigma^2,
\end{aligned}$$

where $\sigma_k^2 = \frac{1}{n}\sum\limits_{i=1}^n \|h_i^k - F_i(x^*)\|^2$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

The lemma above implies that Assumption 2.1 is satisfied with certain parameters.

Proposition G.9 (Proposition 5.4)**.** Let Assumptions 4.1, 4.2, 5.2 hold. Suppose that $\alpha \leq \frac{1}{1+\omega}$. Then, DIANA-SGDA with quantization (15) satisfies Assumption 2.1 with $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|h_i^k - F_i(x^*)\|^2$ and

$$A = \left(\frac{1}{2} + \frac{\omega}{n}\right)\widehat{\ell}, \quad B = \frac{2\omega}{n}, \quad D_1 = \frac{(1+\omega)\sigma^2}{n}, \quad C = \frac{\alpha\widehat{\ell}}{2}, \quad \rho = \alpha, \quad D_2 = \alpha\sigma^2.$$

Proof. To get the result, one needs to apply Assumption 4.1 to estimate $\frac{1}{n}\sum_{i=1}^n \|F_i(x^k) - F_i(x^*)\|^2$ from Lemma G.8. □

G.2.2 Analysis of DIANA-SGDA in the Quasi-Strongly Monotone Case

Applying Theorem 2.2 and Corollary 2.3 with $M = \frac{4\omega}{\alpha n}$, we get the following results.

Theorem G.10. Let F be μ-quasi strongly monotone, Assumptions 4.1, 4.2, 5.2 hold, $\alpha \leq 1/(1+\omega)$, and

$$0 < \gamma \leq \frac{1}{\left(1 + \frac{6\omega}{n}\right)\widehat{\ell}}.$$

Then, for all $k \geq 0$ the iterates produced by DIANA-SGDA satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{\alpha}{2}\right\}\right)^k \mathbb{E}[V_0] + \frac{\gamma^2\sigma^2(1+5\omega)}{n \cdot \min\left\{\gamma\mu, \alpha/2\right\}},$$

where $V_0 = \|x^0 - x^*\|^2 + 4\omega\gamma^2\sigma_0^2/\alpha n$.

Corollary G.11. Let the assumptions of Theorem 5.4 hold. Then, for any $K \geq 0$ one can choose $\alpha = 1/(1+\omega)$ and $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned} &\text{if } K \leq \frac{h}{\mu}, && \gamma_k = \frac{1}{h}, \\ &\text{if } K > \frac{h}{\mu} \text{ and } k < k_0, && \gamma_k = \frac{1}{h}, \\ &\text{if } K > \frac{h}{\mu} \text{ and } k \geq k_0, && \gamma_k = \frac{2}{2h + \mu(k - k_0)}, \end{aligned}$$

where $h = \max\left\{\left(1 + \frac{6\omega}{n}\right)\widehat{\ell}, 2\mu(1+\omega)\right\}$, $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the following inequality holds:

$$\begin{aligned} \mathbb{E}[\|x^K - x^{*,K}\|^2] \quad &\leq \quad 32\max\left\{\frac{\left(1+\frac{6\omega}{n}\right)\widehat{\ell}}{\mu}, 2(1+\omega)\right\} V_0 \exp\left(-\min\left\{\frac{\mu}{\widehat{\ell}(1+\frac{6\omega}{n})}, \frac{1}{1+\omega}\right\} K\right) \\ &\quad + \frac{36(1+5\omega)\sigma^2}{\mu^2 n K}. \end{aligned}$$

G.2.3 Analysis of DIANA-SGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of DIANA-SGDA in the monotone case.

Theorem G.12. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2, 5.2 hold. Assume that

$$0 < \gamma \leq \frac{1}{\left(1 + \frac{4\omega}{n}\right)\widehat{\ell}}.$$

Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by DIANA-SGDA satisfy

$$\begin{aligned} \mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \quad &\leq \quad \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \left(2\widehat{\ell} + \frac{12\omega\widehat{\ell}}{n} + \ell\right)\frac{\|x^0 - x^{*,0}\|^2}{K} \\ &\quad + \left(4 + \gamma\left(2\widehat{\ell} + \frac{12\omega\widehat{\ell}}{n} + \ell\right)\right)\frac{\gamma B\sigma_0^2}{\rho K} \\ &\quad + \gamma\left(\left(2 + \gamma\left(2\widehat{\ell} + \frac{12\omega\widehat{\ell}}{n} + \ell\right)\right)\left(\frac{(1+5\omega)\sigma^2}{n}\right)\right) \\ &\quad + 9\gamma\max_{x^*\in X^*}\|F(x^*)\|^2. \end{aligned}$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary G.13. Let the assumptions of Theorem G.12 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\left(\ell + 2\widehat{\ell} + \frac{12\omega\widehat{\ell}}{n}\right)^{-1}, \frac{\sqrt{\alpha n}}{\sqrt{2\omega\ell\widehat{\ell}}}, \frac{\Omega_{0,\mathcal{C}}}{\sigma\sqrt{K^{(1+3\omega)}/n}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\},$$

This choice of γ implies that $\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right]$ equals

$$\mathcal{O}\left(\frac{(\ell + \widehat{\ell} + \omega\widehat{\ell}/n)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}^2\sqrt{\widehat{\ell}\ell}\sqrt{\omega}}{\sqrt{\alpha n}K} + \frac{\Omega_{0,\mathcal{C}}(\sqrt{(1+\omega)\sigma^2/n} + G_*)}{\sqrt{K}}\right).$$

Proof. The proof follows from the next upper bound $\widehat{\sigma}_0^2$ for σ_0^2 with initialization $h_i^0 = F_i(x^0)$

$$\begin{aligned}
\sigma_0^2 &= \frac{1}{n}\sum_{i=1}^n \|F_i(x^0) - F_i(x^*)\|^2 \\
&\leq \widehat{\ell}\langle F(x^0) - F(x^*), x^0 - x^*\rangle \\
&\leq \widehat{\ell}\|F(x^0) - F(x^*)\| \cdot \|x^0 - x^*\| \\
&\leq \widehat{\ell}\ell\|x^0 - x^*\|^2 \leq \widehat{\ell}\ell\max_{u\in\mathcal{C}}\|x^0 - u\|^2 \leq \widehat{\ell}\ell\Omega_{0,\mathcal{C}}^2.
\end{aligned}$$

Next, applying Corollary D.4 with $\widehat{\sigma}_0 := \sqrt{\widehat{\ell}\ell}\Omega_{0,\mathcal{C}}$, we get the result. □

G.2.4 Analysis of DIANA-SGDA in the Cocoercive Case

Next, using Theorem 2.6, we establish the convergence of DIANA-SGDA in the cocoercive case.

Theorem G.14. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2, 5.2 hold. Assume that

$$0 < \gamma \leq \frac{1}{\left(1 + \frac{4\omega}{n}\right)\widehat{\ell}}.$$

Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by DIANA-SGDA satisfy

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] &\leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \left(3\widehat{\ell} + \frac{18\omega\widehat{\ell}}{n} + 3\ell\right)\frac{\|x^0 - x^{*,0}\|^2}{K} \\
&\quad + \left(6 + \gamma\left(4\widehat{\ell} + \frac{18\omega\widehat{\ell}}{n} + 3\ell\right)\right)\frac{\gamma B\sigma_0^2}{\rho K} \\
&\quad + \gamma\left(\left(3 + \gamma\left(3\widehat{\ell} + \frac{18\omega\widehat{\ell}}{n} + 3\ell\right)\right)\left(\frac{(1+5\omega)\sigma^2}{n}\right)\right).
\end{aligned}$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary G.15. Let the assumptions of Theorem G.14 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\left(3\ell + 3\widehat{\ell} + \frac{18\omega\widehat{\ell}}{n}\right)^{-1}, \frac{\sqrt{\alpha n}}{\sqrt{2\omega\ell\widehat{\ell}}}, \frac{\Omega_{0,\mathcal{C}}}{\sigma\sqrt{K^{(1+3\omega)}/n}}\right\},$$

This choice of γ implies that $\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right]$ equals

$$\mathcal{O}\left(\frac{(\ell+\widehat{\ell}+\omega\widehat{\ell}/n)(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)}{K}+\frac{\Omega_{0,\mathcal{C}}^2\sqrt{\widehat{\ell}\ell}\sqrt{\omega}}{\sqrt{\alpha n}K}+\frac{\Omega_{0,\mathcal{C}}\sqrt{(1+\omega)\sigma^2/n}}{\sqrt{K}}\right).$$

G.3 VR-DIANA-SGDA

In this section, we assume that each F_i has a finite-sum form: $F_i(x)=\frac{1}{m}\sum_{j=1}^{m} F_{ij}(x)$.

Algorithm 6 VR-DIANA-SGDA: VR-DIANA Stochastic Gradient Descent-Ascent Horváth et al. (2019)

Input: starting points $x^0, h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$, $h^0 = \frac{1}{n}\sum\limits_{i=1}^{n} h_i^0$, probability $p \in (0,1]$ stepsizes $\gamma, \alpha > 0$, number of steps K,
for $k = 0$ **to** $K-1$ **do**
  Broadcast x^k to all workers
  for $i = 1, \ldots, n$ in parallel **do**
    Draw a fresh sample j_i^k from the uniform distribution on $[m]$ and compute $g_i^k = F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k) + F_i(w_i^k)$
    $w_i^{k+1} = \begin{cases} x^k, & \text{with probability } p,\\ w_i^k, & \text{with probability } 1-p,\end{cases}$
    $\Delta_i^k = g_i^k - h_i^k$
    Send $\mathcal{Q}(\Delta_i^k)$ to the server
    $h_i^{k+1} = h_i^k + \alpha\mathcal{Q}(\Delta_i^k)$
  end for
  $g^k = h^k + \frac{1}{n}\sum\limits_{i=1}^{n}\mathcal{Q}(\Delta_i^k) = \frac{1}{n}\sum\limits_{i=1}^{n}(h_i^k + \mathcal{Q}(\Delta_i^k))$
  $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
  $h^{k+1} = h^k + \alpha\frac{1}{n}\sum\limits_{i=1}^{n}\mathcal{Q}(\Delta_i^k) = \frac{1}{n}\sum\limits_{i=1}^{n} h_i^k$
end for

G.3.1 Proof of Proposition 5.6

Lemma G.16 (Modification of Lemmas 3 and 7 from Horváth et al. (2019))**.** Let F be ℓ-star-cocoercive and Assumptions 4.1, 4.2, 5.5 hold. Then for all $k \geq 0$ VR-DIANA-SGDA satisfies

$$\begin{aligned}
\mathbb{E}_k\left[g^k\right] &= F(x^k),\\
\mathbb{E}_k\left[\|g^k - F(x^*)\|\right] &\leq \left(\ell + \frac{2\widetilde{\ell}}{n} + \frac{2\omega(\widehat{\ell}+\widetilde{\ell})}{n}\right)\langle F(x^k)-F(x^*), x^k - x^*\rangle + \frac{2(\omega+1)}{n}\sigma_k^2,
\end{aligned}$$

where $\sigma_k^2 = \frac{H^k}{n} + \frac{D^k}{nm}$ with $H^k = \sum\limits_{i=1}^{n}\left\|h_i^k - F_i(x^*)\right\|^2$ and $D^k = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\left\|F_{ij}(w_i^k) - F_{ij}(x^*)\right\|^2$.

Proof. First of all, we derive unbiasedness:

$$\mathbb{E}\left[g^k\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathcal{Q}(g_i^k - h_i^k) + h_i^k\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[g_i^k - h_i^k + h_i^k\right] = \frac{1}{n}\sum_{i=1}^{n}F_i(x^k) = F(x^k).$$

By definition of the variance we get

$$\mathbb{E}_{\mathcal{Q}}\left[\left\|g^k - F(x^*)\right\|^2\right] = \underbrace{\left\|\mathbb{E}_{\mathcal{Q}}\left[g^k\right] - F(x^*)\right\|^2}_{T_1} + \underbrace{\mathbb{E}_{\mathcal{Q}}\left[\left\|g^k - \mathbb{E}_{\mathcal{Q}}\left[g^k\right]\right\|^2\right]}_{T_2}.$$

Next, we derive the upper bounds for terms T_1 and T_2 separately. For T_2 we use unbiasedness of quantization and independence of workers:

$$\begin{aligned}T_2 &= \mathbb{E}_Q\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}(g_i^k - h_i^k) - (g_i^k - h_i^k)\right\|^2\right]\\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{\mathcal{Q}}\left[\left\|\mathcal{Q}(g_i^k - h_i^k) - (g_i^k - h_i^k)\right\|^2\right] \overset{(15)}{\leq} \frac{\omega}{n^2}\sum_{i=1}^n \left\|g_i^k - h_i^k\right\|^2.\end{aligned}$$

Taking $\mathbb{E}_k[\cdot]$ from the both sides of the above inequality, we derive

$$\begin{aligned}\mathbb{E}_k[T_2] &\leq \frac{\omega}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|g_i^k - h_i^k\right\|^2\right] = \frac{\omega}{n^2}\sum_{i=1}^n\left(\left\|\mathbb{E}_k\left[g_i^k - h_i^k\right]\right\|^2 + \mathbb{E}_k\left[\left\|g_i^k - h_i^k - \mathbb{E}_k\left[g_i^k - h_i^k\right]\right\|^2\right]\right)\\ &= \frac{\omega}{n^2}\sum_{i=1}^n\left(\left\|F_i(x^k) - h_i^k\right\|^2 + \mathbb{E}_k\left[\left\|g_i^k - F_i(x^k)\right\|^2\right]\right)\\ &= \frac{\omega}{n^2}\sum_{i=1}^n\left(\left\|F_i(x^k) - h_i^k\right\|^2 + \mathbb{E}_k\left[\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k) - \mathbb{E}_k\left[F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k)\right]\right\|^2\right]\right)\\ &\leq \frac{\omega}{n^2}\sum_{i=1}^n\left(\left\|F_i(x^k) - h_i^k\right\|^2 + \mathbb{E}_k\left[\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k)\right\|^2\right]\right)\\ &\leq \frac{2\omega}{n^2}\sum_{i=1}^n\left(\left\|h_i^k - F_i(x^\star)\right\|^2 + \left\|F_i(x^k) - F_i(x^\star)\right\|^2\right)\\ &\quad + \frac{2\omega}{n^2}\sum_{i=1}^n\left(\mathbb{E}_k\left[\left\|F_{ij_i^k}(w_i^k) - F_{ij_i^k}(x^\star)\right\|^2\right] + \mathbb{E}_k\left[\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(x^\star)\right\|^2\right]\right).\end{aligned}$$

Since j_i^k is sampled uniformly at random from $[m]$, we have

$$\begin{aligned}\mathbb{E}_k[T_2] &\leq \frac{2\omega}{n^2}\sum_{i=1}^n\left(\left\|h_i^k - F_i(x^\star)\right\|^2 + \left\|F_i(x^k) - F_i(x^\star)\right\|^2\right)\\ &\quad + \frac{2\omega}{mn^2}\sum_{i=1}^n\sum_{j=1}^m\left(\mathbb{E}_k\left[\left\|F_{ij}(w_i^k) - F_{ij}(x^\star)\right\|^2\right] + \mathbb{E}_k\left[\left\|F_{ij}(x^k) - F_{ij}(x^\star)\right\|^2\right]\right)\\ &\overset{(14),(19)}{\leq} \frac{2\omega}{n^2}H^k + \frac{2\omega}{mn^2}D^k + \frac{2\omega(\widehat{\ell} + \widetilde{\ell})}{n}\langle F(x^k) - F(x^*), x^k - x^*\rangle.\end{aligned}$$

In last line, we also use the definitions of H^k, D^k. For T_1 we use definition of g^k:

$$T_1 = \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{Q}}\left[\mathcal{Q}(g_i^k - h_i^k) + h_i^k\right] - F(x^*)\right\|^2 = \left\|\frac{1}{n}\sum_{i=1}^n g_i^k - F(x^*).\right\|^2$$

Next, we estimate $\mathbb{E}_k[T_1]$ similarly to $\mathbb{E}_k[T_2]$:

$$\begin{aligned}\mathbb{E}_k[T_1] &= \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k - F(x^*)\right\|^2\right] = \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[g_i^k\right] - F(x^*)\right\|^2 + \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(g_i^k - \mathbb{E}\left[g_i^k\right]\right)\right\|_2^2\right]\\ &= \left\|F(x^k) - F(x^*)\right\|^2 + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|g_i^k - F_i(x^k)\right\|^2\right]\\ &\overset{(4)}{\leq} \ell\langle F(x^k) - F(x^*), x^k - x^*\rangle\\ &\quad + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k) - \mathbb{E}_k\left[F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k)\right]\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&\leq \ell\langle F(x^k) - F(x^*), x^k - x^*\rangle + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_k\left[\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k)\right\|^2\right] \\
&= \ell\langle F(x^k) - F(x^*), x^k - x^*\rangle + \frac{1}{mn^2}\sum_{i=1}^n\sum_{j=1}^m \left\|F_{ij}(x^k) - F_{ij}(w_i^k)\right\|^2 \\
&\leq \ell\langle F(x^k) - F(x^*), x^k - x^*\rangle + \frac{2}{mn^2}\sum_{i=1}^n\sum_{j=1}^m \left(\left\|F_{ij}(w_i^k) - F_{ij}(x^\star)\right\|^2 + \left\|F_{ij}(x^k) - F_{ij}(x^\star)\right\|^2\right) \\
&\overset{(19)}{\leq} \left(\ell + \frac{2\widetilde{\ell}}{n}\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle + \frac{2}{mn^2}D^k.
\end{aligned}
$$

Finally, summing $\mathbb{E}\left[T_1\right]$ and $\mathbb{E}\left[T_2\right]$ we get

$$
\begin{aligned}
\mathbb{E}\left[\left\|g^k - F(x^*)\right\|^2\right] &= \mathbb{E}\left[T_1 + T_2\right] \\
&\leq \left(\ell + \frac{2\widetilde{\ell}}{n}\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle + \frac{2}{mn^2}D^k \\
&\quad + \frac{2\omega}{n^2}H^k + \frac{2\omega}{mn^2}D^k + \frac{2\omega(\widehat{\ell} + \widetilde{\ell})}{n}\langle F(x^k) - F(x^*), x^k - x^*\rangle \\
&\leq \left(\ell + \frac{2\widetilde{\ell}}{n} + \frac{2\omega(\widehat{\ell} + \widetilde{\ell})}{n}\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle + \frac{2\omega}{n^2}H^k + \frac{2(\omega+1)}{mn^2}D^k,
\end{aligned}
$$

which concludes the proof since $\sigma_k^2 = \frac{H^k}{n} + \frac{D^k}{nm}$. □

Lemma G.17 (Modification of Lemmas 5 and 6 from Horváth et al. (2019)). Let F be ℓ-star-cocoercive and Assumptions 4.1, 4.2, 5.5 hold. Suppose that $\alpha \leq \min\left\{\frac{p}{3}; \frac{1}{1+\omega}\right\}$. Then for all $k \geq 0$ VR-DIANA-SGDA satisfies

$$
\mathbb{E}_k\left[\sigma_{k+1}^2\right] \quad \leq \quad (1-\alpha)\sigma_k^2 + \left(p\widetilde{\ell} + 2\alpha(\widetilde{\ell} + \widehat{\ell})\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle,
$$

where $\sigma_k^2 = \frac{H^k}{n} + \frac{D^k}{nm}$ with $H^k = \sum\limits_{i=1}^n \left\|h_i^k - F_i(x^*)\right\|^2$ and $D^k = \sum\limits_{i=1}^n\sum\limits_{j=1}^m \left\|F_{ij}(w_i^k) - F_{ij}(x^*)\right\|^2$.

Proof. We start with considering H^{k+1}:

$$
\begin{aligned}
\mathbb{E}_k\left[H^{k+1}\right] &= \mathbb{E}_k\left[\sum_{i=1}^n \left\|h_i^{k+1} - F_i(x^\star)\right\|^2\right] \\
&= \sum_{i=1}^n \left\|h_i^k - F_i(x^\star)\right\|^2 + \sum_{i=1}^n \mathbb{E}_k\left[2\langle\alpha\mathcal{Q}(g_i^k - h_i^k), h_i^k - F_i(x^\star)\rangle + \alpha^2\left\|\mathcal{Q}(g_i^k - h_i^k)\right\|^2\right] \\
&\overset{(15)}{\leq} H^k + \sum_{i=1}^n \mathbb{E}_k\left[2\alpha\langle g_i^k - h_i^k, h_i^k - F_i(x^\star)\rangle + \alpha^2(\omega+1)\left\|g_i^k - h_i^k\right\|^2\right].
\end{aligned}
$$

Since $\alpha \leq 1/(\omega+1)$, we have

$$
\begin{aligned}
\mathbb{E}_k\left[H^{k+1}\right] &\leq H^k + \mathbb{E}_k\left[\sum_{i=1}^n \alpha\langle g_i^k - h_i^k, g_i^k + h_i^k - 2F_i(x^\star)\rangle\right] \\
&= H^k + \mathbb{E}_k\left[\sum_{i=1}^n \alpha\langle g_i^k - F_i(x^\star) + F_i(x^\star) - h_i^k, g_i^k - F_i(x^\star) + h_i^k - F_i(x^\star)\rangle\right] \\
&= H^k + \mathbb{E}_k\left[\sum_{i=1}^n \alpha\left(\left\|g_i^k - F_i(x^\star)\right\|^2 - \left\|h_i^k - F_i(x^\star)\right\|^2\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= H^k(1-\alpha) + \mathbb{E}_k\left[\sum_{i=1}^n \alpha\left(\left\|g_i^k - F_i(x^\star)\right\|^2\right)\right] \\
&\leq H^k(1-\alpha) + \sum_{i=1}^n \left(2\alpha\mathbb{E}_k\left[\left\|g_i^k - F_i(x^k)\right\|^2\right] + 2\alpha\left\|F_i(x^k) - F_i(x^\star)\right\|^2\right) \\
&= H^k(1-\alpha) + \sum_{i=1}^n \mathbb{E}_k\left[2\alpha\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k) - \mathbb{E}_k\left[F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k)\right]\right\|^2\right] \\
&\quad +2\alpha\sum_{i=1}^n \left\|F_i(x^k) - F_i(x^\star)\right\|^2 \\
&\leq H^k(1-\alpha) + \sum_{i=1}^n \left(\mathbb{E}_k\left[2\alpha\left\|F_{ij_i^k}(x^k) - F_{ij_i^k}(w_i^k)\right\|^2\right] + 2\alpha\left\|F_i(x^k) - F_i(x^\star)\right\|^2\right) \\
&\leq H^k(1-\alpha) + \frac{2\alpha}{m}\sum_{i=1}^n\sum_{j=1}^m \left(\left\|F_{ij}(x^k) - F_{ij}(x^\star)\right\|^2 + \left\|F_{ij}(w_i^k) - F_{ij}(x^\star)\right\|^2\right) \\
&\quad +2\alpha\sum_{i=1}^n \left\|F_i(x^k) - F_i(x^\star)\right\|_2^2 \\
&\overset{(14),(19)}{\leq} H^k(1-\alpha) + \frac{2\alpha}{m}\sum_{i=1}^n\sum_{j=1}^m \left\|F_{ij}(w_{ij}^k) - F_{ij}(x^\star)\right\|_2^2 \\
&\quad +2\alpha n(\widetilde{\ell} + \widehat{\ell})\langle F(x^k) - F(x^*), x^k - x^*\rangle \\
&= H^k(1-\alpha) + \frac{2\alpha}{m}D^k + 2\alpha n(\widetilde{\ell} + \widehat{\ell})\langle F(x^k) - F(x^*), x^k - x^*\rangle.
\end{aligned}
$$

Next, we consider D^{k+1}

$$
\begin{aligned}
\mathbb{E}_k\left[D^{k+1}\right] &= \sum_{i=1}^n\sum_{j=1}^m \mathbb{E}_k\left[\left\|F_{ij}(w_i^{k+1}) - F_{ij}(x^\star)\right\|^2\right] \\
&= \sum_{i=1}^n\sum_{j=1}^m \left[(1-p)\left\|F_{ij}(w_{ij}^k) - F_{ij}(x^\star)\right\|_2^2 + p\left\|F_{ij}(x^k) - F_{ij}(x^\star)\right\|_2^2\right] \\
&\overset{(19)}{\leq} D^k(1-p) + nmp\widetilde{\ell}\langle F(x^k) - F(x^*), x^k - x^*\rangle.
\end{aligned}
$$

It remains put the upper bounds on D^{k+1}, H^{k+1} together and use the definition of σ_{k+1}^2:

$$
\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &= \frac{\mathbb{E}_k\left[H^{k+1}\right]}{n} + \frac{\mathbb{E}_k\left[D^{k+1}\right]}{nm} \\
&\leq (1-\alpha)\frac{H^k}{n} + (1+2\alpha-p)\frac{D^k}{nm} + \left(p\widetilde{\ell} + 2\alpha(\widetilde{\ell} + \widehat{\ell})\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle
\end{aligned}
$$

With $\alpha \leq \frac{p}{3}$ we get $-p \leq -3\alpha$, implying

$$
\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &\leq (1-\alpha)\frac{H^k}{n} + (1-\alpha)\frac{D^k}{nm} + \left(p\widetilde{\ell} + 2\alpha(\widetilde{\ell} + \widehat{\ell})\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle \\
&= (1-\alpha)\sigma_k^2 + \left(p\widetilde{\ell} + 2\alpha(\widetilde{\ell} + \widehat{\ell})\right)\langle F(x^k) - F(x^*), x^k - x^*\rangle.
\end{aligned}
$$

□

The above two lemmas imply that Assumption 2.1 is satisfied with certain parameters.

Proposition G.18 (Proposition 5.6)**.** Let F be ℓ-star-cocoercive and Assumptions 4.1, 4.2, 5.5 hold. Suppose that

$\alpha \leq \min\left\{\frac{p}{3}; \frac{1}{1+\omega}\right\}$. Then, VR-DIANA-SGDA satisfies Assumption 2.1 with

$$A = \left(\frac{\ell}{2} + \frac{\widetilde{\ell}}{n} + \frac{\omega(\widehat{\ell} + \widetilde{\ell})}{n}\right), \quad B = \frac{2(\omega+1)}{n},$$

$$\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - F_i(x^*)\right\|^2 + \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|F_{ij}(w_i^k) - F_{ij}(x^*)\right\|^2,$$

$$C = \left(\frac{p\widetilde{l}}{2} + \alpha(\widetilde{\ell} + \widehat{\ell})\right), \quad \rho = \alpha \leq \min\left\{\frac{p}{3}; \frac{1}{1+\omega}\right\}, \quad D_1 = D_2 = 0.$$

G.3.2 Analysis of VR-DIANA-SGDA in the Quasi-Strongly Monotone Case

Applying Theorem 2.2 and Corollary 2.3 with $M = \frac{4(\omega+1)}{n\alpha}$, we get the following results.

Theorem G.19. Let F be μ-quasi strongly monotone, ℓ-star-cocoercive and Assumptions 4.1, 4.2, 5.5 hold. Suppose that $\alpha \leq \min\left\{\frac{p}{3}; \frac{1}{1+\omega}\right\}$ and

$$0 < \gamma \leq \left(\ell + \frac{10(\omega+1)(\widehat{\ell} + \widetilde{\ell})}{n} + \frac{4(\omega+1)p\widetilde{l}}{\alpha n}\right)^{-1}.$$

Then for all $k \geq 0$ the iterates of VR-DIANA-SGDA satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, {}^{1}\!/_{\alpha n}\right\}\right)^k V_0,$$

where $V_0 = \|x^0 - x^*\|^2 + \frac{4(\omega+1)\gamma^2}{n\alpha}\sigma_0^2$.

Corollary G.20. Let the assumptions of Theorem G.19 hold. Then, for $p = \frac{1}{m}$, $\alpha = \min\left\{\frac{1}{3m}, \frac{1}{1+\omega}\right\}$,

$$\gamma = \left(\ell + \frac{10(\omega+1)(\widehat{\ell} + \widetilde{\ell})}{n} + \frac{4(\omega+1)\max\{3m, 1+\omega\}\widetilde{\ell}}{nm}\right)^{-1}$$

and any $K \geq 0$ we have

$$\mathbb{E}[\|x^k - x^*\|^2] \leq V_0 \exp\left(-\min\left\{\frac{\mu}{\ell + \frac{10(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n} + \frac{4(\omega+1)\max\{3m,1+\omega\}\widetilde{\ell}}{nm}}, \frac{1}{6m}, \frac{1}{2(1+\omega)}\right\}K\right).$$

G.3.3 Analysis of VR-DIANA-SGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of VR-DIANA-SGDA in the monotone case.

Theorem G.21. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2, 5.5 hold. Assume that

$$0 < \gamma \leq \left(\ell + \frac{6(\omega+1)(\widehat{\ell} + \widetilde{\ell})}{n} + \frac{2(\omega+1)p\widetilde{l}}{\alpha n}\right)^{-1}$$

and $\alpha = \min\left\{\frac{p}{3}, \frac{1}{1+\omega}\right\}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by VR-DIANA-SGDA satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^K x^k\right)\right] \leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K}$$

$$+\left(3\ell + \frac{12(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n} + \frac{8(\omega+1)p\widetilde{l}}{\alpha n}\right)\cdot\frac{\|x^0 - x^{*,0}\|^2}{K}$$
$$+\left(4+\gamma\left(3\ell + \frac{12(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n} + \frac{8(\omega+1)p\widetilde{l}}{\alpha n}\right)\right)\frac{\gamma B\sigma_0^2}{\rho K}.$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary G.22. Let the assumptions of Theorem G.21 hold. Then $\forall K > 0$ one can choose $p = \frac{1}{m}$, $\alpha = \min\left\{\frac{1}{3m}, \frac{1}{1+\omega}\right\}$ and γ as

$$\gamma = \min\left\{\frac{1}{3\ell + \frac{12(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n} + \frac{8(\omega+1)\max\{3m,1+\omega\}\widetilde{\ell}}{mn}},\right.$$
$$\left.\frac{\Omega_{0,\mathcal{C}}\sqrt{n}}{\Omega_{0,\mathcal{C}}\sqrt{2\max\{3m,1+\omega\}(\omega+1)(\widetilde{\ell}+\widehat{\ell})\ell}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}.$$

This choice of α and γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{\left(\ell + {}^{(\omega+1)(\widehat{\ell}+\widetilde{\ell})}/_{n} + {}^{(\omega+1)\max\{m,\omega\}\widetilde{\ell}}/_{mn}\right)(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K}\right.$$
$$\left. + \frac{\Omega_{0,\mathcal{C}}^2\sqrt{\max\{m,\omega\}(\omega+1)(\widetilde{\ell}+\widehat{\ell})\ell}}{\sqrt{n}K} + \frac{\Omega_{0,\mathcal{C}}G_*}{\sqrt{K}}\right).$$

Proof. The proof follows from the next upper bound $\widehat{\sigma}_0^2$ for σ_0^2 with initialization $h_i^0 = F_i(x^0)$ and $w_i = x_0$

$$\begin{aligned}
\sigma_0^2 &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|F_{ij}(x^0) - F_{ij}(x^*)\|^2 + \frac{1}{n}\sum_{i=1}^{n}\|F_i(x^0) - F_i(x^*)\|^2 \\
&\le (\widetilde{\ell}+\widehat{\ell})\langle F(x^0) - F(x^*), x^0 - x^*\rangle \\
&\le (\widetilde{\ell}+\widehat{\ell})\|F(x^0) - F(x^*)\|\cdot\|x^0 - x^*\| \\
&\le (\widetilde{\ell}+\widehat{\ell})\ell\|x^0 - x^*\|^2 \le (\widetilde{\ell}+\widehat{\ell})\ell\max_{u\in\mathcal{C}}\|x^0 - u\|^2 \le (\widetilde{\ell}+\widehat{\ell})\ell\Omega_{0,\mathcal{C}}^2.
\end{aligned}$$

Next, applying Corollary D.4 with $\widehat{\sigma}_0 := \sqrt{(\widetilde{\ell}+\widehat{\ell})\ell}\Omega_{0,\mathcal{C}}$, we get the result. □

G.3.4 Analysis of VR-DIANA-SGDA in the Cocoercive Case

Next, using Theorem 2.6, we establish the convergence of VR-DIANA-SGDA in the cocoercive case.

Theorem G.23. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 4.1, 4.2, 5.5 hold. Assume that

$$0 < \gamma \le \left(\ell + \frac{6(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n} + \frac{2(\omega+1)p\widetilde{l}}{\alpha n}\right)^{-1}$$

and $\alpha = \min\left\{\frac{p}{3}, \frac{1}{1+\omega}\right\}$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \ge 0$ the iterates produced by VR-DIANA-SGDA satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \le \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K}$$

$$+\left(6\ell+\frac{18(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n}+\frac{12(\omega+1)p\widetilde{l}}{\alpha n}\right)\cdot\frac{\|x^0-x^{*,0}\|^2}{K}$$
$$+\left(6+\gamma\left(6\ell+\frac{18(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n}+\frac{12(\omega+1)p\widetilde{l}}{\alpha n}\right)\right)\frac{\gamma B\sigma_0^2}{\rho K}.$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary G.24. Let the assumptions of Theorem G.23 hold. Then $\forall K>0$ one can choose $p=\frac{1}{m}$, $\alpha=\min\left\{\frac{1}{3m},\frac{1}{1+\omega}\right\}$ and γ as

$$\gamma = \min\left\{\frac{1}{6\ell+\frac{18(\omega+1)(\widehat{\ell}+\widetilde{\ell})}{n}+\frac{12(\omega+1)\max\{3m,1+\omega\}\widetilde{\ell}}{mn}},\right.$$
$$\left.\frac{\Omega_{0,\mathcal{C}}\sqrt{n}}{\Omega_{0,\mathcal{C}}\sqrt{2\max\{3m,1+\omega\}(\omega+1)(\widetilde{\ell}+\widehat{\ell})\ell}}\right\}.$$

This choice of α and γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K}x^k\right)\right]=\mathcal{O}\left(\frac{\left(\ell+{}^{(\omega+1)(\widehat{\ell}+\widetilde{\ell})}/_{n}+{}^{(\omega+1)\max\{m,\omega\}\widetilde{\ell}}/_{mn}\right)\left(\Omega_{0,\mathcal{C}}^2+\Omega_0^2\right)}{K}\right.$$
$$\left.+\frac{\Omega_{0,\mathcal{C}}^2\sqrt{\max\{m,\omega\}(\omega+1)(\widetilde{\ell}+\widehat{\ell})\ell}}{\sqrt{n}K}\right).$$

G.4 Discussion of the Results in the Monotone and Cocoercive Cases

Beznosikov et al. (2021b) also consider monotone case and derive the following rate for MASHA1 (neglecting the dependence on Lipschitz parameters and the quantities like $\Omega_{0,\mathcal{C}}^2=\max_{u\in\mathcal{C}}\|x^0-u\|^2$): $\mathcal{O}\left(\sqrt{(m+\omega)(1+{}^{\omega}/_{n})}\frac{1}{K}\right)$. In general, due to the term proportional to ${}^{1}/\sqrt{K}$ and due to the relation between (star-)cocoercivity constants and Lipschitz constants our rate $\mathcal{O}\left(\frac{(1+\omega)}{nK}+\frac{(1+\omega)\max\{m,\omega\}}{mnK}+\frac{\sqrt{\max\{m,\omega\}(1+\omega)}}{\sqrt{n}K}+\frac{G_*}{\sqrt{K}}\right)$ our rate is worse than the one from Beznosikov et al. (2021b) (even when $R(x)\equiv 0$, i.e., $G_*=0$). However, when the difference between cocoercivity and Lipschitz constants is not significant, and m,n or ω are sufficiently large, our result in the cocoercive case (Corollary G.24) might be better. Moreover, we emphasize here that Beznosikov et al. (2021b) do not consider SGDA as the basis for their methods. To the best of our knowledge, our results are the first ones for distributed SGDA-type methods with compression derived in the monotone case without assuming (quasi-)strong monotonicity.

H COORDINATE SGDA

In this section, we focus on the coordinate versions of SGDA. To denote i-th component of the vector x we use $[x]_i$. Vectors $e_1, \ldots, e_d \in \mathbb{R}^d$ form a standard basis in $\mathbb{R}^d$.

H.1 CSGDA

Algorithm 7 CSGDA: Coordinate Stochastic Gradient Descent-Ascent

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, number of steps K
2: **for** $k = 0$ **to** $K-1$ **do**
3: Sample uniformly at random $j \in [d]$
4: $g^k = de_j[F(x^k)]_j$
5: $x^{k+1} = \text{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
6: **end for**

H.1.1 CSGDA Fits Assumption 2.1

Proposition H.1. Let F be ℓ-star-cocoercive. Then, CSGDA satisfies Assumption 2.1 with

$$A = d\ell, \quad D_1 = 2d \max_{x* \in X^*}\left[\|F(x^*)\|^2\right], \quad \sigma_k^2 = 0, \quad B = 0, \quad C = 0, \quad \rho = 1, \quad D_2 = 0.$$

Proof. First of all, for all $a \in \mathbb{R}^d$ and for random index j uniformly distributed on $[d]$ we have $\mathbb{E}_j[\|e_j[a]_j\|^2] = \frac{1}{d}\sum_{i=1}^d [a]_j^2 = \frac{1}{d}\|a\|^2$. Using this and $g^k = de_j[F(x^k)]_j$, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &= \mathbb{E}_k\left[\|de_j[F(x^k) - F(x^{*,k})]_j + de_j[F(x^{*,k})]_j - F(x^{*,k})\|^2\right] \\
&\le 2\mathbb{E}_k\left[\|de_j[F(x^k) - F(x^{*,k})]_j\|^2\right] + 2\mathbb{E}_k\left[\|de_j[F(x^{*,k})]_j - F(x^{*,k})\|^2\right] \\
&= 2d\|F(x^k) - F(x^{*,k})\|^2 + 2\mathbb{E}_k\left[\|de_j[F(x^{*,k})]_j - \mathbb{E}_k[de_j[F(x^{*,k})]_j]\|^2\right] \\
&\le 2d\|F(x^k) - F(x^{*,k})\|^2 + 2\mathbb{E}_k\left[\|de_j[F(x^{*,k})]_j\|^2\right] \\
&= 2d\|F(x^k) - F(x^{*,k})\|^2 + 2d\|F(x^{*,k})\|^2.
\end{aligned} \tag{64}$$

Finally, the star-cocoercivity of F implies

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - F(x^{*,k})\|^2\right] &\le 2d\ell\langle F(x^k) - F(x^{*,k}), x^k - x^*\rangle + 2d\|F(x^{*,k})\|^2 \\
&\le 2d\ell\langle F(x^k) - F(x^{*,k}), x^k - x^*\rangle + 2d\max_{x* \in X^*}\left[\|F(x^*)\|^2\right].
\end{aligned}$$

□

H.1.2 Analysis of CSGDA in the Quasi-Strongly Monotone Case

Applying Theorem 2.2 and Corollary 2.3, we get the following results.

Theorem H.2. Let F be μ-quasi strongly monotone and ℓ-star-cocoercive, $0 < \gamma \le 1/2d\ell$. Then for all $k \ge 0$

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \le (1-\gamma\mu)^k\|x^0 - x^{*,0}\|^2 + \frac{2\gamma d}{\mu} \cdot \max_{x* \in X^*}\left[\|F(x^*)\|^2\right].$$

Corollary H.3. Let the assumptions of Theorem H.2 hold. Then, for any $K \ge 0$ one can choose $\{\gamma_k\}_{k \ge 0}$ as follows:

$$\begin{aligned}
\text{if } K \le \frac{2d\ell}{\mu}, \qquad & \gamma_k = \frac{1}{2d\ell}, \\
\text{if } K > \frac{2d\ell}{\mu} \text{ and } k < k_0, \qquad & \gamma_k = \frac{1}{2d\ell},
\end{aligned}$$

$$\text{if } K > \frac{2d\ell}{\mu} \text{ and } k \geq k_0, \qquad \gamma_k = \frac{2}{\mu(4d\ell + \mu(k - k_0))},$$

where $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the following inequality holds:

$$\mathbb{E}[V_K] \leq \frac{64d\ell}{\mu}\|x^0 - x^{*,0}\|^2 \exp\left(-\frac{\mu K}{2d\ell}\right) + \frac{72d}{\mu^2 K} \cdot \max_{x^* \in X^*}\left[\|F(x^*)\|^2\right].$$

H.1.3 Analysis of CSGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of CSGDA in the monotone case.

Theorem H.4. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4 hold. Assume that $\gamma \leq 1/2d\ell$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by CSGDA satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + \frac{5d\ell\|x^0 - x^{*,0}\|^2}{K} + 20\gamma d \cdot \max_{x^* \in X^*}\left[\|F(x^*)\|^2\right].$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary H.5. Let the assumptions of Theorem H.4 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{5d\ell}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{2dK}}\right\}.$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{d\ell(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2) + \ell\Omega_{\mathcal{C}}^2}{K} + \frac{\Omega_{0,\mathcal{C}}G_*}{\sqrt{K}}\right).$$

H.1.4 Analysis of CSGDA in the Cocoercive Case

Next, using Theorem 2.6, we establish the convergence of CSGDA in the cocoercive case.

Theorem H.6. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4 hold. Assume that $\gamma \leq 1/2d\ell$. Then for $\mathrm{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by CSGDA satisfy

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0 - u\|^2\right]}{2\gamma K} + \frac{9d\ell\|x^0 - x^{*,0}\|^2}{K} + 16\gamma d \cdot \max_{x^* \in X^*}\left[\|F(x^*)\|^2\right].$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary H.7. Let the assumptions of Theorem H.6 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{9d\ell}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{2dK}}\right\}.$$

This choice of γ implies

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{d\ell(\Omega_{0,\mathcal{C}}^2 + \Omega_0^2)}{K} + \frac{\Omega_{0,\mathcal{C}}G_*}{\sqrt{K}}\right).$$

H.2 SEGA-SGDA

In this section, we consider a modification of SEGA (Hanzely et al., 2018) – the linearly converging coordinate method for composite optimization problems working even for non-separable regularizers.

Algorithm 8 SEGA-SGDA: SEGA Stochastic Gradient Descent-Ascent Hanzely et al. (2018)

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, number of steps K
2: Set $h^0 = 0$
3: **for** $k = 0$ **to** $K - 1$ **do**
4: Sample uniformly at random $j \in [d]$
5: $h^{k+1} = h^k + e_j([F(x^k)]_j - h^k_j)$
6: $g^k = de_j([F(x^k)]_j - h^k_j) + h^k$
7: $x^{k+1} = \mathrm{prox}_{\gamma R}\left(x^k - \gamma g^k\right)$
8: **end for**

H.2.1 SEGA-SGDA Fits Assumption 2.1

The following result from Hanzely et al. (2018) does not rely on the fact that $F(x)$ is the gradient of some function. Therefore, it holds in our settings as well.

Lemma H.8 (Lemmas A.3 and A.4 from Hanzely et al. (2018))**.** Let Assumption 4.2 hold. Then for all $k \geq 0$ SEGA-SGDA satisfies

$$\begin{aligned}\mathbb{E}_k\left[\|g^k - F(x^*)\|^2\right] &\leq 2d\|F(x^k) - F(x^*)\|^2 + 2d\sigma_k^2,\\ \mathbb{E}_k\left[\sigma_{k+1}^2\right] &\leq \left(1 - \frac{1}{d}\right)\sigma_k^2 + \frac{1}{d}\|F(x^k) - F(x^*)\|^2,\end{aligned}$$

where $\sigma_k^2 = \|h^k - F(x^*)\|^2$.

The lemma above implies that Assumption 2.1 is satisfied with certain parameters.

Proposition H.9. Let F be ℓ-star-cocoercive and Assumption 4.2 holds. Then, SEGA-SGDA satisfies Assumption 2.1 with $\sigma_k^2 = \|h^k - F(x^*)\|^2$ and

$$A = d\ell, \quad B = 2d, \quad D_1 = 0, \quad C = \frac{\ell}{2d}, \quad \rho = \frac{1}{d}, \quad D_2 = 0.$$

Proof. The result follows from Lemma H.8 and star-cocoercivity of F. □

H.2.2 Analysis of SEGA-SGDA in the Quasi-Strongly Monotone Case

Applying Theorem 2.2 and Corollary 2.3 with $M = 4d^2$, we get the following results.

Theorem H.10. Let F be μ-quasi strongly monotone, ℓ-star-cocoercive, Assumption 4.2 holds, and $0 < \gamma \leq \frac{1}{6d\ell}$. Then, for all $k \geq 0$ the iterates produced by SEGA-SGDA satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{1}{2d}\right\}\right)^k \cdot V_0,$$

where $V_0 = \|x^0 - x^*\|^2 + 4d^2\gamma^2\sigma_0^2$.

Corollary H.11. Let the assumptions of Theorem H.10 hold. Then, for $\gamma = \frac{1}{6d\ell}$ and any $K \geq 0$ we have

$$\mathbb{E}[\|x^k - x^*\|^2] \leq V_0 \exp\left(-\min\left\{\frac{\mu}{6d\ell}, \frac{1}{2d}\right\}K\right).$$

H.2.3 Analysis of SEGA-SGDA in the Monotone Case

Next, using Theorem 2.5, we establish the convergence of CSGDA in the monotone case.

Theorem H.12. Let F be monotone, ℓ-star-cocoercive and Assumptions 2.1, 2.4, 4.2 hold. Assume that $\gamma \leq 1/6d\ell$. Then for $\text{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by SEGA-SGDA satisfy

$$\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K} + \frac{8\gamma\ell^2\Omega_{\mathcal{C}}^2}{K} + 13d\ell\cdot\frac{\|x^0-x^{*,0}\|^2}{K}$$
$$+ (4+13\gamma d\ell)\frac{2d\gamma\sigma_0^2}{K} + 9\gamma\cdot\max_{x^*\in X^*}\left[\|F(x^*)\|^2\right].$$

Applying Corollary D.4, we get the rate of convergence to the exact solution.

Corollary H.13. Let the assumptions of Theorem H.12 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{13d\ell}, \frac{\Omega_{0,\mathcal{C}}}{\sqrt{2G^* d}}, \frac{\Omega_{0,\mathcal{C}}}{G_*\sqrt{K}}\right\}.$$

This choice of γ implies

$$\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{d\ell(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)+\ell\Omega_{\mathcal{C}}^2}{K} + \frac{d\Omega_{0,\mathcal{C}}G_*}{K} + \frac{\Omega_{0,\mathcal{C}}G_*}{\sqrt{K}}\right).$$

Proof. The proof follows from the next upper bound $\hat{\sigma}_0^2$ for σ_0^2 with initialization $h_0 = 0$

$$\sigma_0^2 = \|h_0 - F(x^*)\|^2 = \|F(x^*)\|^2 \leq G_*^2.$$

□

H.2.4 Analysis of SEGA-SGDA in the Cocoercive Case

Next, using Theorem D.5, we establish the convergence of CSGDA in the cocoercive case.

Theorem H.14. Let F be ℓ-cocoercive and Assumptions 2.1, 2.4, 4.2 hold. Assume that $\gamma \leq 1/6d\ell$. Then for $\text{Gap}_{\mathcal{C}}(z)$ from (9) and for all $K \geq 0$ the iterates produced by SEGA-SGDA satisfy

$$\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] \leq \frac{3\left[\max_{u\in\mathcal{C}}\|x^0-u\|^2\right]}{2\gamma K} + 21d\ell\cdot\frac{\|x^0-x^{*,0}\|^2}{K}$$
$$+ (6+21\gamma d\ell)\frac{2d\gamma\sigma_0^2}{K}.$$

Applying Corollary D.6, we get the rate of convergence to the exact solution.

Corollary H.15. Let the assumptions of Theorem H.14 hold. Then $\forall K > 0$ one can choose γ as

$$\gamma = \min\left\{\frac{1}{21d\ell}, \frac{\Omega_{0,\mathcal{C}}}{\sqrt{2G^* d}}\right\}.$$

This choice of γ implies

$$\mathbb{E}\left[\text{Gap}_{\mathcal{C}}\left(\frac{1}{K}\sum_{k=1}^{K} x^k\right)\right] = \mathcal{O}\left(\frac{d\ell(\Omega_{0,\mathcal{C}}^2+\Omega_0^2)}{K} + \frac{d\Omega_{0,\mathcal{C}}G_*}{K}\right).$$

H.3 Comparison with Related Work

The summary of rates in the (quasi-) strongly monotone case is provided in Table 3. First of all, our results are the first convergence for solving *regularized* VIPs via coordinate methods. In particular, SEGA-SGDA is the first linearly converging coordinate method for solving regularized VIPs. Next, when $q = 2$ in zoVIA from Sadiev et al. (2021), i.e., Euclidean proximal setup is used, our rate for SEGA-SGDA is better than the one derived for zoVIA in Sadiev et al. (2021) since $\ell \leq L^2/\mu$. Finally, zoscESVIA might have better rate, but it is based on EG and it uses approximation of each component of operator F at each iteration, which makes one iteration of the method costly.

In the monotone and cocoercive cases, our result and the results from Sadiev et al. (2021) are comparable modulo the difference between (star-)cocoercivity and Lipschitz constants.

Table 3: Summary of the complexity results for zeroth-order methods with two-points feedback oracles for solving (1). By complexity we mean the number of oracle calls required for the method to find x such that $\mathbb{E}[\|x - x^*\|^2] \leq \varepsilon$. By default, operator F is assumed to be **μ-strongly monotone** and, as the result, the solution is unique. Our results rely on **μ-quasi strong monotonicity** of F (3). Methods supporting $R(x) \not\equiv 0$ are highlighted with *. Our results are highlighted in green. Notation: $q =$ the parameter depending on the proximal setup, $q = 2$ in Euclidean case and $q = +\infty$ in the ℓ_1-proximal setup; $G_* = \max_{x* \in X^*} \|F(x^*)\|$, which is zero when $R(x) \equiv 0$.

Method	Citation	Assumptions	Complexity
zoscESVIA [(1)]	(Sadiev et al., 2021)	F is L-Lip.[(2)]	$\widetilde{\mathcal{O}}\left(d\frac{L}{\mu}\right)$
zoVIA	(Sadiev et al., 2021)	F is L-Lip.[(2)]	$\widetilde{\mathcal{O}}\left(d^{2/q}\frac{L^2}{\mu^2}\right)$
CSGDA *	**This paper**	F is ℓ-cocoer.	$\widetilde{\mathcal{O}}\left(d\frac{\ell}{\mu} + \frac{dG_*^2}{\mu^2\varepsilon}\right)$
SEGA-SGDA *	**This paper**	F is ℓ-cocoer. As. 4.2	$\widetilde{\mathcal{O}}\left(d + d\frac{\ell}{\mu}\right)$

[(1)] The method is based on Extragradient update rule. Moreover, at each step full operator is approximated.
[(2)] The problem is defined on a bounded set.