Journal of Machine Learning Research 26 (2025) 1-68 Submitted 3/25; Revised 8/25; Published 11/25

VFOSA: Variance-Reduced Fast Operator Splitting Algorithms for Generalized Equations

Quoc Tran-Dinh QUOCTD@EMAIL.UNC.EDU
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill
318 Hanes Hall, CB #3260, NC 27599-3260

Editor: Peter Richtárik

Abstract

We develop two **V**ariance-reduced **F**ast **O**perator **S**plitting **A**lgorithms (VFOSA) to approximate solutions for a class of generalized equations, covering fundamental problems such as minimization, minimax problems, and variational inequalities as special cases. Our approach integrates recent advances in accelerated operator splitting and fixed-point methods, co-hypomonotonicity structure, and variance reduction techniques. First, we introduce a class of variance-reduced estimators and establish their variance-reduction bounds. This class includes both unbiased and biased instances and comprises common estimators as special cases, including SVRG, SAGA, SARAH, and Hybrid-SGD. Second, we design a novel accelerated variance-reduced forward-backward splitting (FBS) method using these estimators to solve generalized equations in both finite-sum and expectation settings. Our algorithm achieves both $\mathcal{O}\big(1/k^2\big)$ and $o\big(1/k^2\big)$ convergence rates on the expected squared norm $\mathbb{E}\big[\|G_\lambda x^k\|^2\big]$ of the FBS residual G_λ, where k is the iteration counter. Additionally, we establish almost sure convergence rates and the almost sure convergence of iterates to a solution of the underlying generalized equation. Unlike existing stochastic operator splitting algorithms, our methods accommodate co-hypomonotone operators, which can include nonmonotone problems arising in recent applications. Third, we specify our method for each concrete estimator mentioned above and derive the corresponding oracle complexity, demonstrating that these variants achieve the best-known oracle complexity bounds without requiring additional enhancement techniques. Fourth, we develop a variance-reduced fast backward-forward splitting (BFS) method, which attains similar convergence results and oracle complexity bounds as our FBS-based algorithm. Finally, we validate our results through numerical experiments and compare their performance with existing methods.

Keywords: Forward-backward splitting method; backward-forward splitting method; Nesterov's acceleration; variance-reduction; co-hypomonotonicity; generalized equation.

1. Introduction

The generalized equation (GE), also known as the nonlinear inclusion, serves as a versatile framework with broad applications across various domains, including operations research, engineering, mechanics, economics, statistics, and machine learning, see, e.g., (Bauschke and Combettes, 2017; Facchinei and Pang, 2003; Phelps, 2009; Burachik and Iusem, 2008; Ryu and Yin, 2022; Ryu and Boyd, 2016). The recent surge in modern machine learning and distributionally robust optimization has reinvigorated interest in minimax problems, which are special cases of GE. These minimax models, particularly in the context of generative

adversary, imitation learning, reinforcement learning, and distributionally robust optimization, can be effectively modeled and solved using the GE framework, see, e.g., (Arjovsky et al., 2017; Faghri et al., 2025; Goodfellow et al., 2014; Kuhn et al., 2025; Madry et al., 2018; Namkoong and Duchi, 2016; Shi et al., 2022; Swamy et al., 2021; Yu et al., 2022). This paper develops two novel classes of stochastic accelerated operator splitting algorithms with variance reduction specifically designed for solving GEs.

(a) **Problem statement.** In this work, we focus on the following generalized equation (also known as an inclusion):

$$\text{Find } x^\star \in \mathbb{R}^p \text{ such that: } 0 \in \Phi x^\star := Fx^\star + Tx^\star, \tag{GE}$$

where $F : \mathbb{R}^p \to \mathbb{R}^p$ is a single-valued mapping, $T : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is a possibly multivalued mapping, and $\Phi := F + T$. We consider two different settings of (GE) as follows.

(F) [**Finite-sum setting**] F is a large finite-sum of the form:

$$Fx = \frac{1}{n}\sum_{i=1}^{n} F_i x, \tag{F}$$

where $F_i : \mathbb{R}^p \to \mathbb{R}^p$ for $i \in [n] := \{1, \cdots, n\}$ and n is often sufficiently large.

(E) [**Expectation setting**] F is equipped with an unbiased stochastic oracle $\mathbf{F}(\cdot, \xi)$, where ξ is a random variable defined on a given probability space $(\Omega, \mathbb{P}, \Sigma)$, i.e., for any $x \in \mathrm{dom}(F)$, we have

$$Fx = \mathbb{E}_\xi\big[\mathbf{F}(x, \xi)\big]. \tag{E}$$

Note that the finite-sum setting (F) can be viewed as a special case of the expectation setting (E) by choosing $\mathbf{F}(x, \xi) = \frac{1}{n\mathbf{p}_i} F_i x$ for $\mathbf{p}_i := \mathbb{P}(\xi = i) \in (0, 1)$. However, since our algorithms have different oracle complexity bounds on each setting, we treat them separately.

The mapping T in (GE) is possibly multivalued and maximally ρ-co-hypomonotone (see Subsection 2.1 for the definition) as stated in Assumption 1.1(iii) below.

(b) **Fundamental assumptions.** To develop our algorithms for solving (GE), we require the following assumptions on (GE).

Assumption 1.1 *The generalized equation* (GE) *satisfies the following conditions:*

(i) $\mathrm{zer}(\Phi) := \{x^\star \in \mathbb{R}^p : 0 \in \Phi x^\star\} \neq \emptyset$ (*i.e., there exists a solution* $x^\star$ *of* (GE)).

(ii) (***Bounded variance***) *For the expectation setting* (E)*, there exists* $\sigma \geq 0$ *such that* $\mathbb{E}_\xi\big[\|\mathbf{F}(x, \xi) - Fx\|^2\big] \leq \sigma^2$ *for all* $x \in \mathrm{dom}(F)$.

(iii) (***Maximal ρ-co-hypomonotonicity***) T *is maximally* ρ*-co-hypomonotone.*

Assumption 1.1(i) and Assumption 1.1(ii) are standard. While Assumption 1.1(i) guarantees that (GE) is solvable, Assumption 1.1(ii) has been widely used in various stochastic methods. It was also modified and generalized in different ways, see, e.g., (Beznosikov et al., 2023; Gorbunov et al., 2020). We use this assumption to derive our oracle complexities in the sequel that depend on σ^2 and a mega batch-size n_k in the expectation setting (E).

Assumption 1.1(iii) includes maximally monotone operators T, but also covers a class of nonmonotone operators as shown in Subsection 2.2 below with concrete examples. If T is maximally monotone, then it already encompasses the normal cone of a nonempty,

closed, and convex set and the subdifferential of a proper, closed, and convex function as special cases. Consequently, under Assumption 1.1, (GE) includes constrained convex minimization and convex-concave saddle-point problems, and monotone [mixed] variational inequality problems (VIPs) as special cases.

Assumption 1.2 *The mapping F in* (GE) *satisfies one of the following assumptions.*

(F) [***The finite-sum setting***] *For the finite-sum setting* (F)*, F is $\frac{1}{L}$-average co-coercive, i.e., for all $x, y \in \mathrm{cl}(\mathrm{dom}(F))$ (the closure of $\mathrm{dom}(F)$), there exists $L > 0$ such that*

$$\langle Fx - Fy, x - y\rangle \geq \tfrac{1}{nL}\textstyle\sum_{i=1}^n \|F_i x - F_i y\|^2. \tag{1}$$

(E) [***The expectation setting***] *For the expectation setting* (E)*, F is $\frac{1}{L}$-co-coercive in expectation, i.e., for all $x, y \in \mathrm{cl}(\mathrm{dom}(F))$, there exists $L > 0$ such that:*

$$\langle Fx - Fy, x - y\rangle = \langle \mathbb{E}_\xi\big[\mathbf{F}(x,\xi) - \mathbf{F}(y,\xi)\big], x - y\rangle \geq \tfrac{1}{L}\mathbb{E}_\xi\big[\|\mathbf{F}(x,\xi) - \mathbf{F}(y,\xi)\|^2\big]. \tag{2}$$

The average co-coercivity (1) is generally stronger than the co-coercivity of F since we have $\frac{1}{n}\sum_{i=1}^n \|F_i x - F_i y\|^2 \geq \|Fx - Fy\|^2$ by Jensen's inequality. Similarly, the co-coercivity in expectation (2) is generally stronger than the co-coercivity of F since $\mathbb{E}_\xi\big[\|\mathbf{F}(x,\xi) - \mathbf{F}(y,\xi)\|^2\big] \geq \|\mathbb{E}_\xi\big[\mathbf{F}(x,\xi) - \mathbf{F}(y,\xi)\big]\|^2 = \|Fx - Fy\|^2$ by Jensen's inequality. Both cases lead to $\langle Fx - Fy, x - y\rangle \geq \frac{1}{L}\|Fx - Fy\|^2$, i.e., F is $\frac{1}{L}$-co-coercive. Consequently, we get $\|Fx - Fy\| \leq L\|x - y\|$, showing that F is L-Lipschitz continuous. See Subsection 2.2 for further discussion and its connection to the gradient of a smooth and convex function.

(c) **Motivating examples.** GE looks simple, but it is sufficiently general to cover various models across disciplines. We recall some important special cases of (GE) here. We also refer to Davis (2022); Peng et al. (2016); Ryu and Boyd (2016) for additional examples.

(i) ***Composite minimization.*** Consider the following composite minimization problem:

$$\min_{x\in\mathbb{R}^p} \big\{\phi(x) := f(x) + g(x)\big\}, \tag{OP}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is convex and L-smooth (i.e., ∇f is L-Lipschitz continuous) and $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is proper and closed, but not necessarily convex.

Let ∇f be the gradient of f and ∂g be the [abstract] subdifferential of g, see (Bauschke et al., 2020). Then, under appropriate regularity assumptions (Rockafellar and Wets, 2004), the optimality condition of (OP) is

$$0 \in \nabla f(x^\star) + \partial g(x^\star). \tag{3}$$

If g is convex, then $x^\star$ solves (3) if and only if it solves (OP). Clearly, (OP) is a special case of (GE) with $F := \nabla f$ and $T := \partial g$. We can also verify Assumptions 1.1 and 1.2 for this special case. For instance, in the finite-sum setting (F), if f is L-average smooth, then ∇f satisfies Assumption 1.2. If g is proper, closed, and convex, then Assumption 1.1(iii) is automatically satisfied with $\rho = 0$.

Problem (OP) covers many representative applications in machine learning and data science (Bottou et al., 2018; Sra et al., 2012; Wright, 2017). As a concrete example, consider the case where $f(x) := \frac{1}{n}\sum_{i=1}^n \ell(\langle Z_i, x\rangle; y_i)$ represents an empirical loss associated with a dataset $\{(Z_i, y_i)\}_{i=1}^n$, and $g(x) := \tau R(x)$ is a regularizer used to promote desirable structures

in the solution x (e.g., sparsity via an ℓ_1-norm or a well-known SCAD regularizer). In this form, (OP) captures many statistical learning problems such as linear regression, logistic regression, and support vector machines; see, e.g., (Friedman et al., 2001).

(ii) ***Minimax problem.*** Consider the following minimax optimization problem:

$$\min_{u\in\mathbb{R}^{p_1}}\max_{v\in\mathbb{R}^{p_2}}\Big\{\mathcal{L}(u,v) := f(u) + \mathcal{H}(u,v) - g^*(v)\Big\}, \tag{MP}$$

where $f : \mathbb{R}^{p_1} \to \mathbb{R}\cup\{+\infty\}$ and $g^* : \mathbb{R}^{p_2} \to \mathbb{R}\cup\{+\infty\}$ are proper, closed, and not necessarily convex, and $\mathcal{H} : \mathbb{R}^{p_1}\times\mathbb{R}^{p_2} \to \mathbb{R}$ is a given jointly differentiable and convex-concave function.

Under appropriate regularity conditions (Bauschke and Combettes, 2017), the optimality condition of (MP) becomes

$$0 \in \begin{bmatrix} \nabla_u\mathcal{H}(u^\star,v^\star) \\ -\nabla_v\mathcal{H}(u^\star,v^\star)\end{bmatrix} + \begin{bmatrix}\partial f(u^\star) \\ \partial g^*(v^\star)\end{bmatrix}. \tag{4}$$

In particular, if f and g^* are convex, then (4) is necessary and sufficient for $(u^\star, v^\star)$ to be an optimal solution of (MP). Otherwise, it is only a necessary condition. If we define $x := [u, v]$, $F := [\nabla_u\mathcal{H}, -\nabla_v\mathcal{H}]$, and $T := [\partial f, \partial g^*]$, then (4) is a special case of (GE). If f and g^* are convex, then T is monotone and automatically satisfies Assumption 1.1(iii). Moreover, under appropriate smoothness conditions on $\mathcal{H}$, Assumption 1.2 also holds.

The minimax problem (MP) serves as a fundamental model in robust and distributionally robust optimization (Ben-Tal et al., 2001; Rahimian and Mehrotra, 2019; Namkoong and Duchi, 2016), two-player games (Ho et al., 2022; Kuhn et al., 1996), fair machine learning (Du et al., 2021; Martinez et al., 2020), and generative adversarial networks (GANs) (Arjovsky et al., 2017; Daskalakis et al., 2018; Goodfellow et al., 2014), among many others.

As a specific example, if $f(u) := \delta_{\Delta_{p_1}}(u)$ and $g^*(v) := \delta_{\Delta_{p_2}}(v)$ are the indicators of the standard simplexes Δ_{p_1} and Δ_{p_2}, respectively, and $\mathcal{H}(u,v) := \langle \mathbf{L}u, v\rangle$ is a bilinear form with a given payoff matrix $\mathbf{L}$, then (MP) reduces to the classical bilinear game problem. Another representative example is a non-probabilistic robust optimization model derived from Wald's minimax framework: $\min_u\big\{\phi(u) := h(u) + f(u) \equiv \max_{v\in\mathcal{V}}\mathcal{H}(u,v) + f(u)\big\}$, where u is a decision variable, v denotes an uncertainty vector over the uncertainty set $\mathcal{V}\subset\mathbb{R}^{p_2}$, and the function $h(u) := \max_{v\in\mathcal{V}}\mathcal{H}(u,v)$ captures the worst-case risk across all possible realizations of v. See (Ben-Tal et al., 2001) for concrete instances.

(iii) ***Variational inequality problems*** (VIPs)**.** If $T = \mathcal{N}_{\mathcal{X}}$, the normal cone of a nonempty, closed, and convex set $\mathcal{X}$ in $\mathbb{R}^p$, then (GE) reduces to

$$\text{Find } x^\star\in\mathcal{X} \text{ such that: } \langle Fx^\star, x - x^\star\rangle \geq 0, \text{ for all } x\in\mathcal{X}. \tag{VIP}$$

More generally, if $T = \partial g$, the subdifferential of a convex function g, then (GE) reduces to a mixed VIP. Since $\mathcal{X}$ is convex, $T = \mathcal{N}_{\mathcal{X}}$ automatically satisfies Assumption 1.1(iii).

The VIP covers many well-known problems in practice, including unconstrained and constrained minimization, minimax problems, complementarity problems, and Nash's equilibria, see also (Facchinei and Pang, 2003; Konnov, 2001) for more details and direct applications in traffic networks and economics.

(iv) ***Fixed-point problem of nonexpansive mapping.*** The fixed-point problem is a fundamental topic in computational mathematics, with numerous applications in numerical analysis, ordinary and partial differential equations, engineering, and physics (Agarwal et al., 2001; Bauschke and Combettes, 2017; Combettes and Pesquet, 2011). Let

$P : \mathbb{R}^p \to \mathbb{R}^p$ be a given nonexpansive mapping, i.e., $\|Px - Py\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^p$. Then, the classical fixed-point problem is stated as follows:

$$\text{Find } x^\star \in \mathbb{R}^p \text{ such that:} \quad x^\star = Px^\star. \tag{FP}$$

This problem is equivalent to (GE) with $F := \mathbb{I} - P$ and $T = 0$, where $\mathbb{I}$ is the identity mapping. It is well-known that P is nonexpansive if and only if F is $(1/2)$-co-coercive. The algorithms developed in this paper for (GE) can be applied to solve (FP). We can also generalize (FP) to a Kakutani's fixed-point problem $x^\star \in Px^\star + Tx^\star$ of a single-valued mapping F and a multivalued mapping T, which is also equivalent to (GE) with $F = P - \mathbb{I}$.

(d) **Motivation and challenges.** Advanced numerical methods such as acceleration, stochastic approximation, and variance reduction, have received significant attention over the past few decades for solving special cases of (GE), including (OP), (MP), and (VIP), due to their broad applications in modern machine learning and data science. Relevant works include, but are not limited to, (Alacaoglu and Malitsky, 2022; Alacaoglu et al., 2021; Davis, 2016, 2022; Defazio et al., 2014; Emmanouilidis et al., 2024; Johnson and Zhang, 2013; Nguyen et al., 2017; Sadiev et al., 2024; Tran-Dinh et al., 2022). Moreover, biased variance-reduced estimators such as SARAH (Nguyen et al., 2017) and Hybrid-SGD (Tran-Dinh et al., 2022) have demonstrated better oracle complexity than their unbiased counterparts, such as SVRG (Johnson and Zhang, 2013) and SAGA (Defazio et al., 2014); see also (Driggs et al., 2020; Pham et al., 2020). However, designing new algorithms for (GE) that combine both acceleration and biased variance-reduction remains a largely unexplored topic. This is due to several challenges, including the following.

(i) First, most convergence analyses of Nesterov's accelerated randomized and stochastic methods for merely convex optimization and convex-concave saddle-point problems rely on the objective function as a key metric for constructing a suitable Lyapunov function. However, such a function does not exist in (GE), and it remains unclear how to define an alternative metric that can play a similar role.

(ii) Second, unlike stochastic methods in optimization, there is a lack of convergence analysis techniques for handling biased estimators in algorithms for solving (GE).

(iii) Third, in convex optimization, accelerated methods achieve faster convergence rates by leveraging convexity (or, equivalently, the monotonicity of [sub]gradients). Extending such an acceleration to settings beyond convexity or monotonicity, particularly in stochastic methods for solving (GE), remains a significant challenge.

Hitherto, most existing methods for (GE) still face these challenges (Driggs et al., 2020, 2022), and only a few works have managed to overcome some of them without compromising convergence guarantees (Cai et al., 2022a, 2024; Condat and Richtárik, 2022). In this paper, we develop new classes of accelerated schemes for solving (GE) that can incorporate both unbiased and biased estimators. This capability enables our methods to achieve better oracle complexity than several existing approaches and cover a broad family of algorithms.

(e) **Our contributions.** Our contributions in this paper consist of the following.

(i) We introduce a class of variance-reduced estimators that includes a broad spectrum of both unbiased and biased variance-reduced instances. We demonstrate that common estimators, including SVRG, SAGA, SARAH, and Hybrid-SGD, fall within this class. Furthermore, we establish the necessary bounds required for our convergence analysis.

(ii) We develop a new class of accelerated forward-backward splitting (FBS) methods with variance reduction to approximate a solution of (GE) in both the finite-sum and expectation settings under appropriate co-coercivity of F and co-hypomonotonicity of T. Our algorithm is single-loop and simple to implement. It also covers a broad range of variance-reduced estimators introduced in (i). Our method achieves both $\mathcal{O}\big(1/k^2\big)$ and $o\big(1/k^2\big)$ convergence rates in expectation on the squared norm of the FBS residual (i.e., $\mathbb{E}[\|G_\lambda x^k\|^2]$), along with several summability bounds. We further prove $o\big(1/k^2\big)$ almost sure convergence rates. We also show that the sequences of iterates generated by our method almost surely converge to a solution of (GE).

(iii) We specify our method to cover four specific estimators: SVRG, SAGA, SARAH, and Hybrid-SGD, each achieving the "best-known"[1] oracle complexity. For the SVRG and SAGA estimators, we establish a complexity of $\widetilde{\mathcal{O}}(n + n^{2/3}\epsilon^{-1})$ in the finite-sum setting (F), and $\widetilde{\mathcal{O}}(\epsilon^{-3})$ in the expectation setting (E). For SARAH and Hybrid-SGD, this complexity improves to $\widetilde{\mathcal{O}}(n + n^{1/2}\epsilon^{-1})$ and $\mathcal{O}\big(\epsilon^{-3}\big)$, respectively.

(iv) Alternatively, we also propose a class of accelerated backward-forward splitting (BFS) algorithms with variance reduction for solving (GE), which attains the same convergence properties and oracle complexities as our accelerated FBS-based method.

Table 1: Comparison of existing variance-reduced single-loop methods and our algorithms

Refs	Ass. on F	Add. Ass.	Estimators	Setting	Residual Rate	Complexity
Work 1	co-coercive	$T=0$, μ-SQM.	SVRG & SAGA	(F)	$\mathcal{O}\big(1/k\big)$	$\mathcal{O}\big((L/\mu)\log(\epsilon^{-1})\big)$
Work 2	monotone	monotone T	SVRG	(F)	$\mathcal{O}\big(1/k\big)$	–
Work 3	co-coercive	monotone T	a class	(F)	$\mathcal{O}\big(1/k^2\big)$	$\mathcal{O}(n + n^{2/3}\epsilon^{-1})$
Work 4	co-coercive	monotone T	SARAH	(F)	$\mathcal{O}\big(1/k^2\big)$	$\widetilde{\mathcal{O}}\big(n + n^{1/2}\epsilon^{-1}\big)$
Work 5	co-coercive	monotone T	SARAH	(E)	–	$\mathcal{O}\big(\epsilon^{-3}\big)$
Ours	co-coercive	co-hypomonotone T	a class	(F)	$\mathcal{O}\big(1/k^2\big)$, $o\big(1/k^2\big)$ $x^k \to x^\star$ *a.s.*	$\widetilde{\mathcal{O}}\big(n + n^{2/3}\epsilon^{-1}\big)$ $\to \widetilde{\mathcal{O}}\big(n + n^{1/2}\epsilon^{-1}\big)$
Ours	co-coercive	co-hypomonotone T	a class	(E)	$\mathcal{O}\big(1/k^2\big)$, $o\big(1/k^2\big)$ $x^k \to x^\star$ *a.s.*	$\widetilde{\mathcal{O}}(\epsilon^{-3}) \to \mathcal{O}(\epsilon^{-3})$

Abbreviations: Refs = References; **Ass.** = Assumptions; **Add. Ass.** = Additional Assumptions; **SQM** = strong quasi-monotonicity; (F) = the finite-sum setting, and (E) = the expectation setting; **Residual rate** = the convergence rate on $\mathbb{E}\big[\|G_\lambda x^k\|^2\big]$, where G_λ is either the equation operator F or the FBS residual in (6); **a class** = a class of variance-reduced estimators satisfying Definition 4; and *a.s.* = almost surely.

References: Work 1 is Davis (2022); **Work 2** is Alacaoglu et al. (2021); Alacaoglu and Malitsky (2022); **Work 3** is Tran-Dinh (2024b); **Work 4** is Cai et al. (2024); and **Work 5** is Cai et al. (2022a).

Table 1 summarizes the most related results to our work. Let us further discuss in detail our contributions and compare our results with the most related works. First, our approach is indirect compared to (Cai et al., 2022a, 2024), i.e., we reformulate (GE) into an equation (or equivalently, a fixed-point problem) before developing our algorithms. This approach offers certain advantages: (i) it enables us to handle a co-hypomonotone operator T with a co-hypomonotonicity modulus ρ that is independent of the algorithmic parameters; and (ii) it enhances the flexibility of our method, making it readily applicable to other reformulations such as FBS and BFS. Second, unlike (Alacaoglu et al., 2021; Alacaoglu and Malitsky, 2022; Bot et al., 2019; Davis, 2022), our new class of variance-reduced estimators is sufficiently broad to cover many existing ones as special cases and can potentially accom-

1. They may be different from some existing results by a poly-logarithmic factor $\log^\nu(n)$ or $\log^\nu(1/\epsilon)$.

modate new estimators. Third, our methods differ from Halpern's fixed-point iterations in (Cai et al., 2022a, 2024), which enables us to employ different parameter update rules than those in Halpern's schemes. This distinction is crucial for achieving faster convergence rates of $o\big(1/k^2\big)$ and allows us to establish both the almost sure convergence of iterates and the $o\big(1/k^2\big)$ almost sure convergence rates. Fourth, our algorithms are accelerated and single-loop, making them easier to implement compared to double-loop or catalyst methods (Khalafi and Boob, 2023; Yang et al., 2020). Fifth, our rates and oracle complexity rely on the metric $\mathbb{E}\big[\|G_\lambda x^k\|^2\big]$. This differs from existing results using a gap or a restricted gap function, which only works for monotone problems. Sixth, our rate offers a $1/k$ factor improvement over non-accelerated methods (Alacaoglu et al., 2021; Alacaoglu and Malitsky, 2022; Davis, 2022). Finally, our oracle complexity matches the best-known results for methods using SARAH-type estimators, without requiring any enhancement strategies such as scheduled restarts or multiple loops, as employed, e.g., in Cai et al. (2022a, 2024).

(f) **Related work.** Problem (GE) and its special cases are well-studied in the literature, see, e.g., (Bauschke and Combettes, 2017; Burachik and Iusem, 2008; Facchinei and Pang, 2003; Phelps, 2009; Ryu and Yin, 2022; Ryu and Boyd, 2016). We focus on the most recent works relevant to our methods in both the finite-sum and expectation settings.

Accelerated methods. Deterministic accelerated methods have been broadly developed to solve (GE) and its special cases in early works (He and Monteiro, 2016; Kolossoski and Monteiro, 2017; Attouch and Peypouquet, 2019), and further studied in subsequent papers (Adly and Attouch, 2021; Attouch and Cabot, 2020; Attouch and Fadili, 2022; Boţ et al., 2024; Chen et al., 2017; Gorbunov et al., 2022b; Kim, 2021; Maingé, 2021; Park and Ryu, 2022; Tran-Dinh, 2024a). These methods are based on Nesterov's acceleration technique (Nesterov, 1983). However, unlike in convex optimization, extending Nesterov's acceleration to monotone inclusions presents a fundamental challenge due to the absence of an objective function, which complicates the construction of a suitable Lyapunov function as mentioned earlier. This limitation necessitates a different approach for solving (GE) (Attouch and Peypouquet, 2019; Maingé, 2021). Our approach builds on insights from (Alcala et al., 2023; Attouch and Peypouquet, 2019; Maingé, 2021; Tran-Dinh, 2024a; Yuan and Zhang, 2024), combined with variance reduction strategies to develop new methods.

Alternatively, Halpern's fixed-point iteration (Halpern, 1967) has recently been proven to achieve a better convergence rates, see (Diakonikolas, 2020; Lieder, 2021; Sabach and Shtern, 2017), matching Nesterov's acceleration schemes. Yoon and Ryu (2021) extended Halpern's method to extragradient-type schemes, relaxing the co-coercivity assumption. Many subsequent works have exploited this idea to other methods, e.g., (Alcala et al., 2023; Cai et al., 2022b; Cai and Zheng, 2023; Lee and Kim, 2021b,a; Park and Ryu, 2022; Tran-Dinh and Luo, 2021; Tran-Dinh, 2023, 2024a). Recently, Tran-Dinh (2024a) established a connection between Nesterov's and Halpern's accelerations for various iterative schemes.

Stochastic methods. Stochastic methods for (GE) and its special cases have been extensively developed; see, e.g., (Juditsky et al., 2011; Kotsalis et al., 2022; Pethick et al., 2023). A number of works exploit mirror-prox and averaging techniques, such as those in (Juditsky et al., 2011; Kotsalis et al., 2022), while others rely on projection or extragradient-type schemes, e.g., (Bohm et al., 2022; Cui and Shanbhag, 2021; Iusem et al., 2017; Kannan and Shanbhag, 2019; Mishchenko et al., 2020; Pethick et al., 2023; Yousefian et al., 2018). Many algorithms employ standard Robbins-Monro's stochastic approximation with fixed or in-

creasing mini-batch sizes. Other works extend the analysis to a broader class of algorithms, including both unbiased and biased estimators, e.g., (Beznosikov et al., 2023; Demidovich et al., 2023; Gorbunov et al., 2022a; Loizou et al., 2021), thereby covering standard stochastic and unbiased variance-reduction methods. The complexity typically depends on an upper bound of the variance, which often leads to inefficient oracle complexity bounds.

Variance-reduction methods. Variance-reduction schemes using control variate techniques are widely developed in optimization, where many estimators have been proposed, including SAGA (Defazio et al., 2014), SVRG (Johnson and Zhang, 2013), SARAH (Nguyen et al., 2017), and Hybrid-SGD (Tran-Dinh et al., 2019). Researchers have adopted these estimators to develop methods for solving (GE). For instance, Davis (2016, 2022) proposed SAGA-type methods for (GE), under a "star" co-coercivity and strong quasi-monotonicity, most relevant to our work. However, we focus on accelerated methods that achieve better convergence rates and complexity. The authors in Alacaoglu et al. (2021); Alacaoglu and Malitsky (2022) employed SVRG estimators to develop variance-reduced extragradient-type methods to solve (VIP), but these are non-accelerated. Other works can be found in Bot et al. (2019); Carmon et al. (2019); Chavdarova et al. (2019); Huang et al. (2022); Palaniappan and Bach (2016); Yu et al. (2022), some of which focus on minimax problems or bilinear matrix games. More recently, Cai et al. (2022a, 2024) exploited Halpern's fixed-point iteration to develop variance-reduction methods, often achieving better oracle complexity by employing the SARAH estimator. All these results differ from ours due to the generalization of Definition 4 and the new accelerated methods that we develop in this paper.

(g) **Paper organization.** The rest of this paper is organized as follows. In Section 2, we recall some related notation, concepts, and technical results used in this paper. We also further discuss our assumptions imposed on (GE). Section 3 introduces a class of variance-reduced estimators and establishes their bounds. Section 4 develops our accelerated forward-backward splitting method with variance-reduction to solve (GE) and establishes its convergence properties. We also specify this algorithm for each concrete estimator to obtain the corresponding variant, and estimate its oracle complexity bound. Section 5 presents an alternative: an accelerated backward-forward splitting method with variance reduction for solving (GE) and establishes similar convergence results as in Section 4. Section 6 provides two numerical examples to validate our results, and compare different methods. For clarity of presentation, all the technical proofs are deferred to the appendix.

2. Background and Mathematical Tools

First, we recall the necessary notation and concepts. Next, we further discuss our Assumptions 1.1 and 1.2. Finally, we prove a key result essential for developing our algorithms.

2.1 Notation and basic concepts

We work with a finite dimensional space $\mathbb{R}^p$ equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and the Euclidean norm $\|\cdot\|$. For a single-valued or a multivalued mapping $T : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, $\mathrm{dom}(T) = \{x \in \mathbb{R}^p : Tx \neq \emptyset\}$ denotes its domain, and $\mathrm{gra}(T) = \{(x,u) \in \mathbb{R}^p \times \mathbb{R}^p : u \in Tx\}$ denotes its graph. In addition, $\mathrm{cl}(\mathrm{dom}(T))$ denotes the closure of $\mathrm{dom}(T)$.

For a convex function f, ∇f denotes its [sub]gradient, and ∂f denotes its [abstract] subdifferential. For a given symmetric matrix $\mathbf{X}$, $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ denote its largest and smallest eigenvalues, respectively. We also use standard $\mathcal{O}(\cdot)$ and $o(\cdot)$ for convergence rates and complexity bounds, and $\widetilde{\mathcal{O}}(s)$ for $\mathcal{O}(s\log^{\nu}(s))$ (hiding a poly-log factor).

Let $\mathcal{F}_k$ be the σ-algebra generated by all the randomness arising from the algorithm, including $x^0, x^1, \cdots, x^k$, up to the current iteration k. Let $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$ denote the conditional expectation given $\mathcal{F}_k$, and let $\mathbb{E}[\cdot]$ denote the total expectation.

Next, let us recall the concepts of co-hypomonotonicity, monotonicity, and co-coercivity for operators, see, e.g., (Bauschke and Combettes, 2017; Bauschke et al., 2020) for details.

Definition 1 *For a multivalued mapping $T : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, we say that:*

- *T is ρ-co-hypomonotone if there exists $\rho \geq 0$ such that*

$$\langle u - v, x - y\rangle \geq -\rho\|u - v\|^2, \quad \text{for all } (x,u),(y,v) \in \mathrm{gra}(T). \tag{5}$$

  Here, ρ is referred to as the co-hypomonotonicity modulus of T.
- *T is monotone if (5) holds with $\rho = 0$, i.e., $\langle u-v, x-y\rangle \geq 0$ for $(x,u),(y,v) \in \mathrm{gra}(T)$.*
- *T is maximally [co-hypo]monotone if $\mathrm{gra}(T)$ is not properly contained in the graph of any other [co-hypo]monotone mapping.*

If T is single-valued, then (5) reduces to $\langle Tx - Ty, x - y\rangle \geq -\rho\|Tx - Ty\|^2$ for all $x, y \in \mathrm{dom}(T)$. A co-hypomonotone mapping is not necessarily monotone, see Subsection 2.2 for concrete examples. The co-hypomonotonicity concept in Definition 1 is global. We say that T is locally ρ-co-hypomonotone around $(\bar{x}, \bar{u}) \in \mathrm{gra}(T)$ if there exists a neighborhood $\mathcal{W}$ of $(\bar{x}, \bar{u})$ such that for all $(x,u),(y,v) \in \mathrm{gra}(T) \cap \mathcal{W}$, we have $\langle u - v, x - y\rangle \geq -\rho\|u - v\|^2$.

Definition 2 *Given a single-valued mapping $F : \mathbb{R}^p \to \mathbb{R}^p$, we say that:*

- *F is $\frac{1}{L}$-co-coercive if there exists $L > 0$ such that*

$$\langle Fx - Fy, x - y\rangle \geq \tfrac{1}{L}\|Fx - Fy\|^2, \quad \text{for all } x, y \in \mathrm{dom}(F).$$

- *F is L-Lipschitz continuous if $\|Fx - Fy\| \leq L\|x - y\|$ for all $x, y \in \mathrm{dom}(F)$, where $L \geq 0$ is the Lipschitz constant. In particular, if $L = 1$, then F is nonexpansive.*

If F is $\frac{1}{L}$-co-coercive, then it is also monotone and L-Lipschitz continuous.

The operator $J_T x := \{w \in \mathbb{R}^p : x \in w + Tw\}$ is called the resolvent of T, often denoted by $J_T x = (\mathbb{I} + T)^{-1}x$, where $\mathbb{I}$ is the identity mapping. If T is monotone, then J_T is singled-valued, and if T is maximally monotone, then J_T is singled-valued and $\mathrm{dom}(J_T) = \mathbb{R}^p$.

2.2 Further discussion of Assumptions 1.1 and 1.2

(a) **Pros and cons of Assumption 1.2.** Similar to variance-reduction methods using control variate techniques in optimization, we require Assumption 1.2 in our methods. This assumption has some pros and cons as follows.

(i) ***Pros***. First, if $Fx = \nabla f(x)$, the gradient of a differentiable convex function, then the $\frac{1}{L}$-co-coercivity of F is equivalent to the convexity and L-smoothness of f (i.e., ∇f is L-Lipschitz continuous). Therefore, Assumption 1.2 covers convex and L-smooth functions as special cases, including the finite-sum and expectation settings.

Second, the $\frac{1}{L}$-co-coercivity of F is equivalent to the nonexpansiveness of $G = \mathbb{I} - \frac{2}{L}F$, see (Bauschke and Combettes, 2017, Proposition 4.11). Therefore, our methods can also be applied to find approximate fixed-points of a nonexpansive operator. Note that several

problems without satisfying Assumption 1.2 can be reformulated equivalently to a fixed-point problem of a nonexpansive operator, and thus can be indirectly solved by our methods. More specifically, as shown in Tran-Dinh (2024a); Tran-Dinh and Luo (2025), there are several ways to reformulate (GE) and its special cases into a co-coercive equation (e.g., using the Moreau-Yosida approximation, Douglas-Rachford splitting, or three-operator splitting techniques). This approach possibly expands the applicability of our methods to other problem classes, including monotone inclusions and VIPs.

(ii) ***Cons***. Though Assumption 1.2 is reasonable and relatively broad, it may have some extreme cases. For instance, it does not directly cover general linear mappings $Fx := \mathbb{F}x+q$ for a given square matrix $\mathbb{F}$ and a vector q, unless $\mathbb{F}$ is positive definite. One way to handle this extreme case is to consider its Moreau-Yosida's approximation instead of F itself.

(b) **Examples of co-hypomonotone operators.** We provide here two examples of co-hypomonotone operators. However, other examples exist, see, e.g., (Evens et al., 2023).

Example 1. Consider $Tx := \mathbb{T}x + s$, where $\mathbb{T}$ is symmetric and invertible, but not positive semidefinite, and $s \in \mathbb{R}^p$ is given. Assume that $\rho := -\lambda_{\min}(\mathbb{T}^{-1}) \geq 0$. Then, T is ρ-co-hypomonotone and thus satisfies Assumption 1.1(iii). Generally, it is easy to check that if $\mathbb{T} + \mathbb{T}^\top + 2\rho\mathbb{T}^\top\mathbb{T} \succeq 0$ for some $\rho \geq 0$, then T is ρ-co-hypomonotone.

Next, we consider $Fx := \mathbb{F}x + q$, where $\mathbb{F}$ is a symmetric positive semidefinite matrix and $q \in \mathbb{R}^p$. Clearly, F satisfies Assumption 1.2 with $L := \lambda_{\max}(\mathbb{F})$. However, $\Phi := F + T$ is nonmonotone if we impose $\lambda_{\min}(\mathbb{F} + \mathbb{T}) < 0$. In addition, it is possible to choose $\mathbb{F}$ and $\mathbb{T}$ such that $L\rho < 1$, which satisfies the range condition of $L\rho$ in Lemma 3 below.

Example 2. Let $\Psi : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}$ be a twice continuously differentiable and μ-convex-concave saddle function for some $\mu \in \mathbb{R}$, i.e., $\nabla^2_{uu}\Psi(x) \succeq \mu\mathbb{I}$ and $-\nabla^2_{vv}\Psi(x) \succeq \mu\mathbb{I}$ for all $x = [u, v] \in \mathbb{R}^{p_1+p_2}$. We say that Ψ is α-interaction dominant, see (Grimmer et al., 2023), if there exist some $\alpha > -\frac{1}{\rho}$ and $\rho \in \big(0, \frac{1}{\max\{-\mu,0\}}\big)$ such that for all $x \in \mathbb{R}^{p_1+p_2}$, we have

$$\begin{aligned}\nabla^2_{uu}\Psi(x) + \nabla^2_{uv}\Psi(x)(\tfrac{1}{\rho}\mathbb{I} - \nabla^2_{vv}\Psi(x))^{-1}\nabla^2_{vu}\Psi(x) \quad &\succeq \alpha\mathbb{I},\\ -\nabla^2_{vv}\Psi(x) + \nabla^2_{vu}\Psi(x)(\tfrac{1}{\rho}\mathbb{I} + \nabla^2_{uu}\Psi(x))^{-1}\nabla^2_{uv}\Psi(x) &\succeq \alpha\mathbb{I}.\end{aligned}$$

As proven in Evens et al. (2023, Proposition 4.17), if $\alpha > 0$, then the saddle mapping $Tx := [\nabla_u\Psi(x), -\nabla_v\Psi(x)]$ is ρ-co-hypomonotone. The α-interaction dominance notion was studied in Grimmer et al. (2023) for nonconvex-nonconcave minimax problems.

2.3 Equivalent reformulations

The first step of our approach is to reformulate (GE) into an equation using either the forward-backward splitting (FBS) residual or the backward-forward splitting (BFS) residual mapping. These reformulations are also equivalent to the fixed-point problem (FP).

(a) **Forward-backward splitting reformulation.** We consider the following **forward-backward splitting** residual mapping of (GE):

$$G_\lambda x := \tfrac{1}{\lambda}\big(x - J_{\lambda T}(x - \lambda Fx)\big), \tag{6}$$

for a given $\lambda > 0$, and $J_{\lambda T}$ is the resolvent of λT. Then, $x^\star$ **solves** (GE) **iff** $G_\lambda x^\star = 0$, i.e.:

$$0 \in \Phi x^\star = Fx^\star + Tx^\star \quad \Leftrightarrow \quad G_\lambda x^\star = 0. \tag{7}$$

The equation $G_\lambda x^\star = 0$ can also be rewritten equivalently to the fixed-point formulation: $x^\star = J_{\lambda T}(x^\star - \lambda F x^\star)$ of the FBS operator $J_{\lambda T}((\cdot) - \lambda F(\cdot))$.

(b) **Backward-forward splitting reformulation.** Alternatively, we can also consider the following **backward-forward splitting** residual mapping (Attouch et al., 2018) of (GE):

$$S_\lambda u := F(J_{\lambda T} u) + \tfrac{1}{\lambda}(u - J_{\lambda T} u), \tag{8}$$

for a given $\lambda > 0$. Then, $x^\star$ **solves** (GE) **iff** $u^\star$ **solves** $S_\lambda u^\star = 0$ **and** $x^\star = J_{\lambda T} u^\star$, i.e.:

$$0 \in \Phi x^\star = F x^\star + T x^\star \quad \Leftrightarrow \quad S_\lambda u^\star = 0 \quad \text{and} \quad x^\star = J_{\lambda T} u^\star. \tag{9}$$

(c) **Comparison between** (6) **and** (8)**.** Note that G_λ in (8) does not preserve the finite-sum or the expectation structure similar to F due to the composition $J_{\lambda T} \circ (\mathbb{I} - \lambda F)$. In contrast, S_λ in (8) maintains this structure of F. Moreover, the primal variable of G_λ is x, which directly corresponds to the variable x in (GE). However, the primal variable of S_λ is u, which is indirectly related to x via the resolvent $x = J_{\lambda T} u$, making x a *shadow* variable.

To develop our algorithms, we require further properties of G_λ and S_λ as stated in the following lemma, whose proof can be found in Appendix A.2.

Lemma 3 *For* (GE)*, suppose that F is $\frac{1}{L}$-co-coercive and T is maximally ρ-co-hypomonotone such that $L\rho < 1$. For given $\hat{L}$ and λ such that $\hat{L} \geq L$, $\hat{L}\rho < 1$, and $\rho < \lambda \leq \frac{2(1+\sqrt{1-\hat{L}\rho})}{\hat{L}}$, we define $\bar{\beta} := \frac{\lambda(4-\hat{L}\lambda)-4\rho}{4(1-\rho\hat{L})} \geq 0$ and $\Lambda := \frac{\hat{L}-L}{L\hat{L}} \geq 0$. Then*

(i) *if G_λ is defined by* (6)*, then for all $x, y \in \mathrm{dom}(F+T)$, we have*

$$\langle G_\lambda x - G_\lambda y, x - y\rangle \geq \bar{\beta}\|G_\lambda x - G_\lambda y\|^2 + \Lambda L \langle Fx - Fy, x - y\rangle. \tag{10}$$

(ii) *if S_λ is defined by* (8)*, then for all $u, v \in \mathbb{R}^p$, we have*

$$\langle S_\lambda u - S_\lambda v, u - v\rangle \geq \bar{\beta}\|S_\lambda u - S_\lambda v\|^2 + \Lambda L \langle F(J_{\lambda T} u) - F(J_{\lambda T} v), J_{\lambda T} u - J_{\lambda T} v\rangle. \tag{11}$$

(iii) *if additionally $\lambda \geq 2\rho$, then $\|J_{\lambda T} x - J_{\lambda T} y\| \leq \|x - y\|$ for all $x, y \in \mathrm{dom}(J_{\lambda T})$, i.e., the resolvent $J_{\lambda T}$ is nonexpansive.*

Unlike deterministic methods that only require the co-coercivity of G_λ and S_λ, which were already proven in the literature, see, e.g., (Bauschke and Combettes, 2017), we need the new bounds (10) and (11) for our analysis, which allow us to cover a ρ-co-hypomonotone operator T in (GE). This mapping is not necessarily monotone as demonstrated earlier.

3. A Class of Variance-Reduced Estimators

We are given an unbiased stochastic oracle $\mathbf{F}(\cdot, \xi)$ of F such that for any $x \in \mathrm{dom}(F)$ we have $Fx = \mathbb{E}_\xi\big[\mathbf{F}(x, \xi)\big]$ in both settings (F) and (E). We denote by $\widetilde{F}(x, \mathcal{S})$ an unbiased stochastic estimator of Fx constructed from an i.i.d. sample $\mathcal{S}$ of ξ by querying $\mathbf{F}(\cdot, \xi)$.

3.1 The class of variance-reduced stochastic estimators

Given a sequence of iterates $\{x^k\}$ generated by our algorithm, we consider the following class of variance-reduced estimators $\widetilde{F}^k$ of Fx^k that covers various instances specified later.

Definition 4 *Let $\{x^k\}_{k\geq 0}$ be generated by an algorithm and $\widetilde{F}^k := \widetilde{F}(x^k, \mathcal{S}_k)$ be a stochastic estimator of Fx^k constructed from an i.i.d. sample $\mathcal{S}_k$ adapted to a filtration $\{\mathcal{F}_k\}$. We say that $\widetilde{F}^k$ satisfies a $\boldsymbol{VR}(\Delta_k; \kappa_k, \Theta_k, \sigma_k)$ (variance-reduction) property if there exist three parameters $\kappa_k \in (0, 1]$, $\Theta_k \geq 0$, and $\sigma_k \geq 0$, and a random variable $\Delta_k \geq 0$ such that*

$$\begin{aligned} \mathbb{E}\big[\|\widetilde{F}^k - Fx^k\|^2 \mid \mathcal{F}_k\big] &\leq \mathbb{E}\big[\Delta_k \mid \mathcal{F}_k\big], \\ \mathbb{E}\big[\Delta_k \mid \mathcal{F}_k\big] &\leq (1-\kappa_k)\Delta_{k-1} + \Theta_k\bar{\mathcal{E}}_k + \sigma_k^2, \end{aligned} \tag{12}$$

almost surely for $k \geq 0$, where $\bar{\mathcal{E}}_k := \mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big]$, $x^{-1} := x^0$, and $\Delta_{-1} := 0$.

Note that $\widetilde{F}^k$ covers both unbiased and biased estimators of Fx^k since we do not impose the unbiased condition $\mathbb{E}_{\mathcal{S}_k}\big[\widetilde{F}^k\big] = Fx^k$. In the finite-sum setting (F), $\bar{\mathcal{E}}_k$ in Definition 4 reduces to $\bar{\mathcal{E}}_k := \frac{1}{n}\sum_{i=1}^n \|F_i x^k - F_i x^{k-1}\|^2$. Moreover, from (12), we also have $\mathbb{E}\big[\Delta_0\big] \leq \sigma_0^2$.

We highlight that Definition 4 is different from existing works, including (Driggs et al., 2020), as we only require the condition (12), making it broader to cover several existing estimators that may violate the definition in Driggs et al. (2020). It is also broader than and different from the class of unbiased estimators in Tran-Dinh (2024b) and Tran-Dinh (2025). Our class is also different from other general classes of estimators such as (Beznosikov et al., 2023; Gorbunov et al., 2022a; Loizou et al., 2021) for stochastic optimization algorithms since they require the unbiasedness and additional or different conditions.

3.2 Concrete variance-reduced estimators

In this subsection we present four different variance-reduced estimators satisfying Definition 4: L-SVRG, SAGA, L-SARAH, and Hybrid-SGD, which will be used to develop methods for solving (GE) in this paper. Note that the L-SVRG and SAGA are unbiased, while the L-SARAH and Hybrid-SGD are biased. Though we only focus on these four estimators, we believe that other variance-reduced estimators such as SAG (Le Roux et al., 2012; Schmidt et al., 2017), SARGE (Driggs et al., 2022), SEGA (Hanzely et al., 2018), and JacSketch (Gower et al., 2021) can possibly be used in our methods.

3.2.1 Loopless-SVRG estimator

The SVRG estimator was introduced in Johnson and Zhang (2013), and its loopless version was proposed in Kovalev et al. (2020). We show that this estimator satisfies Definition 4.

For given Fx in (GE), its unbiased stochastic oracle $\mathbf{F}(\cdot,\xi)$, two iterates x^k and $\tilde{x}^k$, and an i.i.d. sample $\mathcal{S}_k$ of size b_k, we construct

$$\widetilde{F}^k := \bar{F}\tilde{x}^k + \mathbf{F}(x^k, \mathcal{S}_k) - \mathbf{F}(\tilde{x}^k, \mathcal{S}_k), \tag{L-SVRG}$$

where, for $k \geq 1$, $\tilde{x}^k$ is updated by

$$\tilde{x}^k := \begin{cases} x^{k-1} & \text{with probability } \mathbf{p}_k \\ \tilde{x}^{k-1} & \text{with probability } 1 - \mathbf{p}_k. \end{cases} \tag{13}$$

Here, $\mathbf{p}_k \in [\underline{\mathbf{p}}, 1)$ is a given probability, $\tilde{x}^0 := x^0$, and $\mathbf{F}(\cdot, \mathcal{S}_k) := \frac{1}{b_k}\sum_{\xi_i \in \mathcal{S}_k} \mathbf{F}(\cdot, \xi_i)$ is a mini-batch estimator of F. The quantity $\bar{F}\tilde{x}^k$ is constructed by one of the two options:

- **Full-batch evaluation.** We choose $\bar{F}\tilde{x}^k := F\tilde{x}^k$.

- **Mega-batch evaluation.** We compute $\bar{F}\tilde{x}^k := \frac{1}{n_k}\sum_{\xi\in\bar{\mathcal{S}}_k}\mathbf{F}(\tilde{x}^k,\xi)$, an unbiased estimator of $F\tilde{x}^k$ using a mega-batch $\bar{\mathcal{S}}_k$ of size n_k. Then, we have $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\bar{F}\tilde{x}^k\big] = F\tilde{x}^k$ and $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\|\bar{F}\tilde{x}^k - F\tilde{x}^k\|^2\big] \le \frac{\sigma^2}{n_k}$, where σ^2 is given in Assumption 1.1(ii) and $n_k \ge n_{k-1}$.

Note that the full batch evaluation $\bar{F}\tilde{x}^k := F\tilde{x}^k$ is often computed for the finite-sum setting (F), while one can use a mega-batch evaluation for the expectation setting (E).

The following lemma shows that $\widetilde{F}^k$ satisfies Definition 4, whose proof is in Appendix B.1.

Lemma 5 *Let $\widetilde{F}^k$ be constructed by* (L-SVRG) *and $\hat{\Delta}_k := \frac{1}{b_k}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi)-\mathbf{F}(\tilde{x}^k,\xi)\|^2\big]$ for $b_k \ge b_{k-1}$. Then, for any $\tau > 0$, we have*

$$\mathbb{E}_k\big[\|\widetilde{F}^k - Fx^k\|^2\big] \le (1+\tau)\mathbb{E}_k\big[\hat{\Delta}_k\big] + \tfrac{1+\tau}{\tau}\mathbb{E}_k\big[\|\bar{F}\tilde{x}^k - F\tilde{x}^k\|^2\big]. \tag{14}$$

For any $\alpha \in (0,1)$ and $\bar{\mathcal{E}}_k$ defined in Definition 4, we have

$$\mathbb{E}_k\big[\hat{\Delta}_k\big] \le (1-\alpha\mathbf{p}_k)\,\hat{\Delta}_{k-1} + \tfrac{1}{(1-\alpha)b_k\mathbf{p}_k}\cdot\bar{\mathcal{E}}_k. \tag{15}$$

Consequently, $\widetilde{F}^k$ satisfies the $\boldsymbol{VR}(\Delta_k;\kappa_k,\Theta_k,\sigma_k)$ property in Definition 4 with $\Delta_k := 2\hat{\Delta}_k + \frac{2\sigma^2}{\tau n_k}$, $\kappa_k := \alpha\mathbf{p}_k$, $\Theta_k := \frac{2}{(1-\alpha)b_k\mathbf{p}_k}$, and $\sigma_k^2 := \frac{2\alpha\mathbf{p}_k\sigma^2}{n_k}$.

In particular, if we choose $\bar{F}\tilde{x}^k := F\tilde{x}^k$, then $\widetilde{F}^k$ satisfies the $\boldsymbol{VR}(\Delta_k;\kappa_k,\Theta_k,\sigma_k)$ property in Definition 4 with $\Delta_k := \hat{\Delta}_k$, $\kappa_k := \alpha\mathbf{p}_k$, $\Theta_k := \frac{1}{(1-\alpha)b_k\mathbf{p}_k}$, and $\sigma_k^2 := 0$.

3.2.2 SAGA estimator for finite-sum setting (F)

The SAGA estimator was introduced in Defazio et al. (2014) for finite-sum convex minimization. We apply it to the finite-sum setting (F). It is constructed as follows.

For a given F defined in the finite-sum setting (F) of (GE), a given sequence $\{x^k\}_{k\ge 0}$, and a given i.i.d. sample $\mathcal{S}_k$ of size b_k, for $k \ge 1$, we update $\hat{F}_i^k$ for all $i \in [n]$ as

$$\hat{F}_i^k := \begin{cases} F_i x^{k-1} & \text{if } i \in \mathcal{S}_k, \\ \hat{F}_i^{k-1} & \text{if } i \notin \mathcal{S}_k. \end{cases} \tag{16}$$

Then, we construct a SAGA estimator for Fx^k as follows:

$$\widetilde{F}^k := \tfrac{1}{n}\textstyle\sum_{i=1}^n \hat{F}_i^k + F_{\mathcal{S}_k}x^k - \hat{F}_{\mathcal{S}_k}^k, \tag{SAGA}$$

where $F_{\mathcal{S}_k}x^k := \frac{1}{b_k}\sum_{i\in\mathcal{S}_k}F_i x^k$ and $\hat{F}_{\mathcal{S}_k}^k := \frac{1}{b_k}\sum_{i\in\mathcal{S}_k}\hat{F}_i^k$. Moreover, we need to store n component $\hat{F}_i^k$ computed so far for all $i \in [n]$ in a table $\mathbb{T}_k := [\hat{F}_1^k, \hat{F}_2^k, \cdots, \hat{F}_n^k]$ initialized at $\hat{F}_i^0 := F_i x^0$ for all $i \in [n]$. SAGA requires significant memory to store $\mathbb{T}_k$ if n and p are both large. We have the following result, whose proof is given in Appendix B.2.

Lemma 6 *Let $\widetilde{F}^k$ be constructed by* (SAGA) *such that $b_{k-1} - \frac{(1-\alpha)b_k b_{k-1}}{2n} \le b_k \le b_{k-1}$ for all $k \ge 1$ and a given $\alpha \in (0,1)$, and $\Delta_k := \frac{1}{nb_k}\sum_{i=1}^n\|F_i x^k - \hat{F}_i^k\|^2$. Then, we have*

$$\begin{cases} \mathbb{E}_k\big[\|\widetilde{F}^k - Fx^k\|^2\big] &\le \mathbb{E}_k\big[\Delta_k\big], \\ \mathbb{E}_k\big[\Delta_k\big] &\le \big(1-\frac{\alpha b_k}{n}\big)\Delta_{k-1} + \frac{\Theta_k}{n}\sum_{i=1}^n\|F_i x^k - F_i x^{k-1}\|^2, \end{cases} \tag{17}$$

where $\Theta_k := \frac{(3-\alpha)n}{(1-\alpha)b_k^2}$.

Consequently, $\widetilde{F}^k$ satisfies the $\boldsymbol{VR}(\Delta_k;\kappa_k,\Theta_k,\sigma_k)$ property in Definition 4 with $\kappa_k := \frac{\alpha b_k}{n} \in (0,1]$, Δ_k and Θ_k given above, and $\sigma_k^2 = 0$.

3.2.3 Loopless-SARAH estimator

The SARAH estimator was introduced in Nguyen et al. (2017) for finite-sum convex optimization. Its loopless variant (L-SARAH) was studied in Driggs et al. (2020); Li et al. (2019, 2020), and recently in Cai et al. (2022a, 2024). It is constructed as follows.

Given $\{x^k\}$ and an i.i.d. sample $\mathcal{S}_k$ of size b_k, we construct an estimator $\widetilde{F}^k$ of Fx^k as

$$\widetilde{F}^k := \begin{cases} \widetilde{F}^{k-1} + \mathbf{F}(x^k, \mathcal{S}_k) - \mathbf{F}(x^{k-1}, \mathcal{S}_k) & \text{with probability } 1 - \mathbf{p}_k, \\ \bar{F}x^k & \text{with probability } \mathbf{p}_k, \end{cases} \tag{L-SARAH}$$

where $\mathbf{F}(\cdot, \mathcal{S}_k) := \frac{1}{b_k}\sum_{\xi_i \in \mathcal{S}_k} \mathbf{F}(\cdot, \xi_i)$, $\widetilde{F}^0 := \bar{F}x^0$, and $\mathbf{p}_k \in [\underline{\mathbf{p}}, 1)$ is a given probability of the switching rule in L-SARAH. The estimator $\bar{F}x^k$ is constructed by one of the two options:

- **Full-batch evaluation.** We compute $\bar{F}x^k := Fx^k$.
- **Mega-batch evaluation.** We compute $\bar{F}x^k := \frac{1}{n_k}\sum_{\xi \in \bar{\mathcal{S}}_k} \mathbf{F}(x^k, \xi)$ using a mega-batch $\bar{\mathcal{S}}_k$ of size n_k. In this case, we again have $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\bar{F}x^k\big] = Fx^k$ and $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\|\bar{F}x^k - Fx^k\|^2\big] \le \frac{\sigma^2}{n_k}$, where σ^2 is given in Assumption 1.1(ii).

The following lemma shows that $\widetilde{F}^k$ satisfies Definition 4, whose proof is in Appendix B.3.

Lemma 7 *Let $\widetilde{F}_k$ be constructed by* (L-SARAH) *and $\bar{\mathcal{E}}_k$ be defined in Definition 4. Then*

$$\mathbb{E}_k\big[\|\widetilde{F}^k - Fx^k\|^2\big] \le (1 - \mathbf{p}_k)\|\widetilde{F}^{k-1} - Fx^{k-1}\|^2 + \mathbf{p}_k\|\bar{F}x^k - Fx^k\|^2 + \tfrac{1-\mathbf{p}_k}{b_k}\cdot\bar{\mathcal{E}}_k. \tag{18}$$

Consequently, $\widetilde{F}^k$ satisfies the $\boldsymbol{VR}(\Delta_k; \kappa_k, \Theta_k, \sigma_k)$ property in Definition 4 with $\Delta_k := \|\widetilde{F}^k - Fx^k\|^2$, $\kappa_k = \mathbf{p}_k$, $\Theta_k := \frac{1}{b_k}$, and $\sigma_k^2 := \frac{\mathbf{p}_k\sigma^2}{n_k}$. In particular, if we choose $\bar{F}x^k := Fx^k$, then $\widetilde{F}^k$ satisfies Definition 4 with the same Δ_k, κ_k, and Θ_k, but with $\sigma_k = 0$.

3.2.4 Hybrid SGD estimator

The hybrid stochastic gradient estimator (HSGD) was introduced in Tran-Dinh et al. (2019, 2022) to construct biased variance-reduced estimators for nonconvex optimization. We extend it here for operator F to solve (GE). It is constructed as follows.

Given $\{x^k\}$ generated by our algorithm, an initial estimate $\widetilde{F}^0$ such that $\mathbb{E}_0\big[\|\widetilde{F}^0 - Fx^0\|^2\big] \le \frac{\sigma^2}{n_0}$, and an i.i.d. sample $\mathcal{S}_k$ of size b_k, we construct $\widetilde{F}^k$ for Fx^k as follows:

$$\widetilde{F}^k := (1 - \tau_k)\big[\widetilde{F}^{k-1} + \mathbf{F}(x^k, \mathcal{S}_k) - \mathbf{F}(x^{k-1}, \mathcal{S}_k)\big] + \tau_k\bar{F}x^k, \tag{HSGD}$$

where $\tau_k \in [0, 1]$ is a given weight, $\bar{F}x^k$ is an unbiased estimator of Fx^k constructed from an i.i.d. sample $\hat{\mathcal{S}}_k$ of size $\hat{b}_k$, i.e., $\mathbb{E}_{\hat{\mathcal{S}}_k}\big[\bar{F}x^k\big] = Fx^k$ and $\mathbb{E}_{\hat{\mathcal{S}}_k}\big[\|\bar{F}x^k - Fx^k\|^2\big] \le \frac{\sigma^2}{\hat{b}_k}$ for $k \ge 1$. Here, we allow $\mathcal{S}_k$ and $\hat{\mathcal{S}}_k$ to be dependent or even identical. Our HSGD estimator covers the following special cases.

- If $\tau_k = 0$, then it reduces to the SARAH estimator (Nguyen et al., 2017).
- If $\tau_k = 1$, then $\widetilde{F}^k = \bar{F}x^k$ as a mini-batch unbiased estimator.
- If $\bar{F}x^k = \mathbf{F}(x^k, \mathcal{S}_k)$ (i.e., $\mathcal{S}_k \equiv \hat{\mathcal{S}}_k$), then $\widetilde{F}_k$ reduces to the STORM estimator developed independently and concurrently in Cutkosky and Orabona (2019).

The following lemma provides a key property of (HSGD), whose proof is in Appendix B.4.

Lemma 8 *Let $\widetilde{F}^k$ be constructed by* (HSGD), $\Delta_k := \|\widetilde{F}^k - Fx^k\|^2$, $\delta_k^2 := \mathbb{E}_{\hat{\mathcal{S}}_k}\big[\|\bar{F}x^k - Fx^k\|^2\big]$, *and $\bar{\mathcal{E}}_k$ be defined in Definition 4. Then, the following statements hold.*

(i) *If $\mathcal{S}_k$ is independent of $\hat{\mathcal{S}}_k$, then*

$$\mathbb{E}_k\big[\Delta_k\big] \le (1-\tau_k)^2\Delta_{k-1} + \tfrac{(1-\tau_k)^2}{b_k}\bar{\mathcal{E}}_k + \tau_k^2\delta_k^2. \tag{19}$$

(ii) *If $\mathcal{S}_k$ and $\hat{\mathcal{S}}_k$ are dependent or identical, then*

$$\mathbb{E}_k\big[\Delta_k\big] \le (1-\tau_k)^2\Delta_{k-1} + \tfrac{2(1-\tau_k)^2}{b_k}\bar{\mathcal{E}}_k + 2\tau_k^2\delta_k^2. \tag{20}$$

(iii) *The estimator $\widetilde{F}^k$ satisfies the $\boldsymbol{VR}(\Delta_k;\kappa_k,\Theta_k,\sigma_k)$ property in Definition 4 with Δ_k given above, $\kappa_k := 1-(1-\tau_k)^2$, and*

- $\Theta_k = \frac{(1-\tau_k)^2}{b_k}$ *and* $\sigma_k^2 = \frac{\tau_k^2\sigma^2}{\hat{b}_k}$, *if $\mathcal{S}_k$ is independent of $\hat{\mathcal{S}}_k$;*
- $\Theta_k = \frac{2(1-\tau_k)^2}{b_k}$ *and* $\sigma_k^2 = \frac{2\tau_k^2\sigma^2}{\hat{b}_k}$, *otherwise.*

4. Variance-Reduced Accelerated Forward-Backward Splitting Method

In this section, we propose a novel **V**ariance-reduced **F**ast **O**perator **S**plitting (forward-backward splitting) **A**lgorithm to solve (GE), abbreviated by (VFOSA$_+$). Instead of using a specific stochastic estimator as in many existing works, like (Cai et al., 2024; Davis, 2022), we develop an algorithmic framework that covers all estimators satisfying Definition 4.

4.1 Variance-reduced fast FBS algorithmic framework

(a) **The proposed algorithm.** Motivated by Nesterov's acceleration techniques (Nesterov, 1983, 2004), our VFOSA$_+$ framework for solving (GE) is presented as follows: *Starting from $x^0 \in \mathrm{dom}(\Phi)$, we set $z^0 := x^0$, and at each iteration $k \ge 0$, we update*

$$\begin{cases} y^k &:= \frac{t_k-1}{t_k}x^k + \frac{1}{t_k}z^k, \\ x^{k+1} &:= y^k - \eta_k\widetilde{G}_\lambda^k, \\ z^{k+1} &:= z^k + \nu(x^{k+1}-y^k), \end{cases} \tag{VFOSA$_+$}$$

where $t_k > 0$, $\eta_k > 0$, and $\nu \in (0,1]$ are given parameters, determined later. Here, $\widetilde{G}_\lambda^k$ is a stochastic estimator of $G_\lambda x^k$ defined by (6), *which is constructed as follows:*

$$\widetilde{G}_\lambda^k := \tfrac{1}{\lambda}\big(x^k - J_{\lambda T}(x^k - \lambda\widetilde{F}^k)\big), \tag{21}$$

where $\widetilde{F}^k$ is a stochastic estimator of Fx^k satisfying Definition 4.

By the non-expansiveness of $J_{\lambda T}$ stated in Lemma 3, one can easily show that

$$\|\widetilde{G}_\lambda^k - G_\lambda x^k\| \le \|\widetilde{F}^k - Fx^k\|. \tag{22}$$

This relation shows that if $\widetilde{F}^k$ well approximates Fx^k, then $\widetilde{G}_\lambda^k$ well approximates $G_\lambda x^k$.

(b) **The implementation version.** By combining (VFOSA$_+$), (21), and the update rules in Theorem 12, we obtain Algorithm 1, which is presented for implementation purposes.

Algorithm 1 is single-loop, and at each iteration k, it requires one evaluation $\widetilde{F}^k$ of Fx^k and one evaluation $J_{\lambda T}$ of T, while other steps are only scalar-vector multiplications or vector additions. For simplicity of implementation, we can use the following parameters:

Algorithm 1 (**V**ariance-reduced **F**ast [FB] **O**perator **S**plitting **A**lgorithm (**VFOSA**$_+$))

1: **Initialization:** Take an initial point $x^0 \in \mathrm{dom}(\Phi)$.
2: Choose μ, λ, and β from Theorem 12. Set $\nu := \frac{\mu}{2}$, $r := 2 + \frac{1}{\mu}$, and $z^0 := x^0$.
3: **For** $k := 0, \cdots, k_{\max}$ **do**
4: Construct an estimator $\widetilde{F}^k$ of Fx^k satisfying Definition 4.
5: Update $t_k := \mu(k + r)$ and $\eta_k := \frac{2\beta(t_k - 1)}{t_k - \nu}$.
6: Update the following steps:

$$\begin{cases} y^k &:= \frac{t_k - 1}{t_k}x^k + \frac{1}{t_k}z^k, \\ w^k &:= J_{\lambda T}(x^k - \lambda\widetilde{F}^k), \\ x^{k+1} &:= y^k - \frac{\eta_k}{\lambda}(x^k - w^k), \\ z^{k+1} &:= z^k + \nu(x^{k+1} - y^k). \end{cases}$$

7: **End For**

- Choose $\mu := 0.95 \cdot \frac{2}{3}$ and $r := 5$ (but other values of r such as $r = 10$ still work).
- Given an estimate of L, choose $\hat{L} := L + \zeta$ and $\lambda := \frac{1}{L+\zeta}$ for some small $\zeta > 0$.
- Given $\rho \geq 0$, compute $\bar{\beta} := \frac{\lambda(4 - \hat{L}\lambda) - 4\rho}{4(1 - \rho\hat{L})}$ and set $\beta := \frac{(2-\mu)\bar{\beta}}{2+\mu}$. In particular, if $\rho = 0$ (i.e., T is monotone), then we can choose $\lambda := \frac{1}{\hat{L}}$ and $\bar{\beta} := \frac{\lambda(4 - \hat{L}\lambda)}{4}$.

This parameter configuration reflects what we mainly used for our experiments in Section 6. However, in practice, depending on applications, we can find appropriate parameters which possibly improve the performance of Algorithm 1, while still still satisfying Theorem 12.

(c) **Comparison to Nesterov's acceleration in convex optimization.** Suppose that we apply (VFOSA$_+$) to solve the composite convex minimization problem (OP), where f is convex and L-smooth and g is proper, closed, and convex. In this case, Assumptions 1.1(iii) and 1.2 automatically hold. The key step is $x^{k+1} := y^k - \eta_k\widetilde{G}^k_\lambda$, which becomes $x^{k+1} := y^k - \frac{\eta_k}{\lambda}(x^k - \mathrm{prox}_{\lambda g}(x^k - \lambda\widetilde{\nabla} f(x^k))$. This is a proximal-gradient step using the gradient mapping $\mathcal{G}_\lambda(x) := \frac{1}{\lambda}(x - \mathrm{prox}_{\lambda g}(x - \lambda\nabla f(x)))$. Thus, (VFOSA$_+$) reduces to

$$\begin{cases} y^k &:= \big(1 - \frac{1}{t_k}\big)x^k + \frac{1}{t_k}z^k, \\ x^{k+1} &:= y^k - \frac{\eta_k}{\lambda}(x^k - \mathrm{prox}_{\lambda g}(x^k - \lambda\widetilde{\nabla} f(x^k)), \\ z^{k+1} &:= z^k + \nu(x^{k+1} - y^k), \end{cases} \tag{23}$$

where $\widetilde{\nabla} f(x^k)$ is a stochastic estimator of $\nabla f(x^k)$. This scheme is a new algorithm for solving the convex optimization problem (OP).

Note that the proximal-gradient variant of Nesterov's accelerated methods (Nesterov, 1983) is equivalent to FISTA in (Beck and Teboulle, 2009) for solving (OP), which can be expressed as follows using the gradient mapping value $\mathcal{G}_\lambda(y^k)$ and $\lambda = \eta_k$:

$$\begin{cases} y^k &:= \big(1 - \frac{1}{t_k}\big)x^k + \frac{1}{t_k}z^k, \\ x^{k+1} &:= y^k - \frac{\eta_k}{\lambda}(y^k - \mathrm{prox}_{\lambda g}(y^k - \lambda\nabla f(y^k)) \equiv \mathrm{prox}_{\lambda g}(y^k - \lambda\nabla f(y^k)), \\ z^{k+1} &:= z^k + t_k(x^{k+1} - y^k). \end{cases} \tag{24}$$

Clearly, our scheme (23) has some similarity to (24) in terms of structure, and y^k and x^{k+1} updates. However, there are two key differences between (23) and (24).

- First, (23) evaluates the gradient mapping $\mathcal{G}_\lambda$ at x^k instead of at y^k as in (24).
- Second, (23) uses a fixed parameter $\nu \in (0,1)$ in the last step instead of t_k as in (24).

These two differences fundamentally change the convergence analysis of our method.

Due to the second and third lines, VFOSA$_+$ is also different from the accelerated schemes in (Attouch and Cabot, 2020; Kim, 2021; Maingé, 2022, 2021; Tran-Dinh, 2024a) since these methods were though derived from Nesterov's accelerated techniques (Nesterov, 1983), they were aided by a different "gradient" correction or "Hessian-driven" damping term.

4.2 Key estimates for convergence analysis

(a) **Lyapunov function.** To analyze the convergence of (VFOSA$_+$), for given $\Lambda \geq 0$ and $\bar{\beta} \geq 0$ in Lemma 3, we construct the following functions w.r.t. the iteration counter k:

$$\begin{aligned}
\mathcal{L}_k &:= \beta a_k \|G_\lambda x^k\|^2 + t_{k-1}\langle G_\lambda x^k, x^k - z^k\rangle + \tfrac{c_k}{2\nu\beta}\|z^k - x^\star\|^2,\\
\mathcal{Q}_k &:= \mathcal{L}_k + \big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2\\
&\quad + \Lambda L t_{k-1}(t_{k-1}-1)\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle,\\
\mathcal{P}_k &:= \mathcal{Q}_k + \tfrac{[\mu(1-\kappa_k)\Gamma_k + \beta]t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1},
\end{aligned} \tag{25}$$

where $\beta > 0$, $\mu \in (0,1]$, and $\nu \in [0,1]$ are given in (VFOSA$_+$), and $s > 0$ and $\Gamma_k \geq 0$ are given parameters, determined later. The quantities Δ_k and κ_k are given in Definition 4, and the coefficients a_k and c_k are respectively given by

$$a_k := t_{k-1}[t_{k-1} - 1 - s(1-\nu)] \quad \text{and} \quad c_k := \tfrac{(1-\mu)[(t_k-\nu)(t_{k-1}-1)+\mu(1-\nu)]}{2(t_{k-1}-1)(t_k-1)}. \tag{26}$$

(b) **Key lemmas.** The following three lemmas provide key bounds for our analysis. We first state the first important lemma, whose proof can be found in Appendix C.1.

Lemma 9 *Suppose that Assumptions 1.1 and 1.2 hold for* (GE). *Let $\mathcal{L}_k$ be defined by* (25) *and $\{(x^k, y^k, z^k)\}$ be generated by* (VFOSA$_+$) *using t_k and η_k respectively as*

$$t_k := \mu(k+r) \quad \text{and} \quad \eta_k = \tfrac{2\beta(t_k-1)}{t_k-\nu}, \tag{27}$$

for given $\mu \in (0,1]$, $r \geq 0$, and $\beta > 0$. Then, for any $s > 0$, we have

$$\begin{aligned}
\mathcal{L}_k - \mathcal{L}_{k+1} \geq\ & \beta\varphi_k\|G_\lambda x^k\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}\\
&+ \big[\bar{\beta} - (1+s)\beta\big]t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 - \psi_k\|e^k\|^2,
\end{aligned} \tag{28}$$

where $\bar{\beta}$ and Λ are given constants in Lemma 3, $e^k := \widetilde{F}^k - Fx^k$, $\mathcal{E}_{k+1} := L\langle Fx^{k+1} - Fx^k, x^{k+1} - x^k\rangle$, and the coefficients φ_k and ψ_k are respectively given by

$$\begin{aligned}
\varphi_k &:= t_{k-1}\big[t_{k-1} - 1 - s(1-\nu)\big] - \tfrac{(t_k-1)}{t_k-\nu}\big[t_k(t_k - 2 + \nu - s(1-\nu)) + 2(1-\mu)\nu\big],\\
\psi_k &:= \beta(t_k-1)\big[\tfrac{t_k(t_k-1)}{s(t_k-\nu)} + \tfrac{2\nu(1-\mu)}{t_k-\nu} + \tfrac{(1-\mu)\nu(t_{k-1}-1)}{\mu(1-\nu)}\big].
\end{aligned} \tag{29}$$

Lemma 9 is of independent interest as it can serve as a core step to analyze inexact variants of (VFOSA$_+$) beyond this work. For example, we can use it to analyze inexact methods (either deterministic or stochastic), where the approximation error $e^k := \widetilde{F}^k - Fx^k$ between $\widetilde{F}^k$ and Fx^k is adaptively controlled along the iterations. For instance, we can assume that $\mathbb{E}_k\big[\|e^k\|^2\big] \le \delta_0 L^2\|x^k - x^{k-1}\|^2$ for a given factor $\delta_0 \ge 0$.

Next, we show that $\mathcal{L}_k$ is lower bounded in Lemma 10, whose proof is in Appendix C.2.

Lemma 10 *Under the same setting as in Lemma 9, $\mathcal{L}_k$ defined by (25) satisfies*

$$\mathcal{L}_k \ge \tfrac{A_k}{2}\|G_\lambda x^k\|^2 + \tfrac{(t_{k-1}-1)[(1-\mu-2\nu)t_k+\nu(1+\mu)]+\mu(1-\mu)(1-\nu)}{4\nu\beta(t_{k-1}-1)(t_k-1)}\|z^k - x^\star\|^2, \tag{30}$$

where $A_k := \beta t_{k-1}[t_{k-1} - 2 - 2s(1-\nu)] + 2\bar{\beta}t_{k-1}$. Moreover, $\mathcal{Q}_k$ and $\mathcal{P}_k$ defined by (25) satisfy $\mathcal{P}_k \ge \mathcal{Q}_k \ge \mathcal{L}_k \ge 0$.

Finally, Lemma 11 states a descent property of $\mathcal{P}_k$, whose proof is in Appendix C.3.

Lemma 11 *Under the same setting as in Lemma 9 and $\lambda \ge 2\rho$, suppose further that κ_k in Definition 4, Γ_k in (25), and ψ_k in (29) satisfy the following condition:*

$$2\psi_k + \big[\Gamma_{k+1}(1-\kappa_{k+1}) + \tfrac{\beta}{\mu}\big]t_k(t_k-1) \le \Gamma_k t_{k-1}(t_{k-1}-1). \tag{31}$$

Then, for $\bar{\mathcal{E}}_k$ defined in Definition 4 and $\mathcal{P}_k$ defined by (25), we have

$$\begin{aligned}\mathcal{P}_k \;\ge\; & \mathbb{E}_k\big[\mathcal{P}_{k+1}\big] + \beta\varphi_k\|G_\lambda x^k\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle \\ & + \big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 \\ & - \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2 + \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1} + \tfrac{(2\Lambda - \Gamma_k\Theta_k)t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k.\end{aligned} \tag{32}$$

In the proofs of Lemmas 9, 10, and 11, we use Young's inequality several times. We do not attempt to optimize its coefficients, resulting in somewhat loose bounds on our results.

4.3 Convergence analysis of VFOSA$_+$

Now, utilizing the above three technical lemmas, we are ready to state and prove our convergence results in the following theorem, whose proof is given in Appendix C.4

Theorem 12 *Suppose that Assumptions 1.1 and 1.2 hold for (GE) such that $L\rho < 1$. Let $\{(x^k, y^k, z^k)\}$ be generated by (VFOSA$_+$) using an estimator $\widetilde{F}^k$ for Fx^k satisfying Definition 4. Let $\hat{L} > L$ be such that $\hat{L}\rho < 1$, and $\bar{\beta}$ and Λ be defined in Lemma 3. Assume that we choose λ, μ, r, ν, and β, and update t_k and η_k as follows:*

$$\begin{aligned}& 2\rho \le \lambda < \tfrac{2(1+\sqrt{1-\hat{L}\rho})}{\hat{L}}, \quad 0 < \mu < \tfrac{2}{3}, \quad r \ge 2 + \tfrac{1}{\mu}, \quad \nu := \tfrac{\mu}{2}, \\ & 0 < \beta \le \tfrac{(2-\mu)\bar{\beta}}{2+\mu}, \quad t_k := \mu(k+r), \quad \text{and} \quad \eta_k := \tfrac{2\beta(t_k-1)}{t_k-\nu}.\end{aligned} \tag{33}$$

Suppose further that, for all $k \ge 0$, κ_k and Θ_k in Definition 4 and Γ_k in (25) satisfy

$$\kappa_k \ge 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_{k-1}(t_{k-1}-1)} + \tfrac{5\beta}{\mu\Gamma_k} \quad \text{and} \quad \Gamma_k\Theta_k \le 2\Lambda. \tag{34}$$

Then, for all $K \geq 0$, G_λ defined by (6) satisfies

$$\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \leq \frac{2(\Psi_0^2 + E_0^2 + B_{K-1})}{\mu^2(K+r-1)^2}, \tag{35}$$

where Ψ_0^2, E_0^2, and B_K are respectively given by

$$\begin{aligned} \Psi_0^2 &:= \mu^2 r^2 \mathbb{E}\big[\|G_\lambda x^0\|^2\big] + \tfrac{2r-1}{4\beta^2(\mu r-1)}\|x^0 - x^\star\|^2, \\ E_0^2 &:= \tfrac{\mu r^2(\Gamma_0\mu+\beta)}{2\beta}\sigma_0^2, \\ B_K &:= \tfrac{\Lambda}{\beta}\textstyle\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k}. \end{aligned} \tag{36}$$

In addition, for any $K \geq 0$, we also have

$$\begin{aligned} \textstyle\sum_{k=0}^K \big[(2-3\mu)\mu(k+r) + 6\mu^2\big]\mathbb{E}\big[\|G_\lambda x^k\|^2\big] &\leq 2\Psi_0^2 + 2E_0^2 + 2B_K, \\ \textstyle\sum_{k=0}^K (k+r)(k+r-\mu^{-1})\mathbb{E}\big[\|\widetilde{F}^k - Fx^k\|^2\big] &\leq \tfrac{2(\Psi_0^2+E_0^2+B_K)}{\beta\mu}, \\ \textstyle\sum_{k=0}^K (k+r)(k+r-\mu^{-1})\mathbb{E}\big[\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\big] &\leq \tfrac{(2-\mu)\beta(\Psi_0^2+E_0^2+B_K)}{\mu^2[(2-\mu)\bar{\beta}-(2+\mu)\beta]}. \end{aligned} \tag{37}$$

Here, the last summability bound in (37) requires $0 < \beta < \frac{(2-\mu)\bar{\beta}}{2+\mu}$.

Note that the conclusions of Theorem 12 hold under the condition (34), and the right-hand side bounds depend on the quantity B_K. Next, we state both $\mathcal{O}\big(1/k^2\big)$ and $o\big(1/k^2\big)$ convergence rates of (VFOSA$_+$) under the following condition:

$$B_\infty := \frac{\Lambda}{\beta}\sum_{k=0}^\infty t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} < +\infty. \tag{38}$$

The proof of the following theorem is deferred to Appendix C.5.

Theorem 13 *Under the same conditions and settings as in Theorem 12, and assuming that the condition (38) also holds, for all $k \geq 0$, we have*

$$\mathbb{E}\big[\|G_\lambda x^k\|^2\big] \leq \frac{2(\Psi_0^2 + E_0^2 + B_\infty)}{\mu^2(k+r-1)^2}. \tag{39}$$

In addition, we have the following summability bounds:

$$\begin{aligned} \textstyle\sum_{k=0}^\infty (k+1)\mathbb{E}\big[\|G_\lambda x^k\|^2\big] &< +\infty, \\ \textstyle\sum_{k=0}^\infty (k+1)^2\mathbb{E}\big[\|\widetilde{F}^k - Fx^k\|^2\big] &< +\infty, \\ \textstyle\sum_{k=0}^\infty (k+1)\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] &< +\infty. \end{aligned} \tag{40}$$

We also have the following $o\big(1/k^2\big)$ convergence rates in expectation:

$$\begin{aligned} \lim_{k\to\infty} k^2\mathbb{E}\big[\|x^{k+1} - x^k\|^2\big] &= 0, \\ \lim_{k\to\infty} k^2\mathbb{E}\big[\|G_\lambda x^k\|^2\big] &= 0. \end{aligned} \tag{41}$$

Finally, we state the following almost sure convergence properties of (VFOSA$_+$). The proof of this theorem can be found in Appendix C.6.

Theorem 14 *Under the same conditions and settings as in Theorem 12, and assuming that* (38) *holds,* $0 < \beta < \frac{(2-\mu)\bar{\beta}}{2+\mu}$, *and there exist* $\underline{\Theta} > 0$ *and* $\underline{\Gamma} > 0$ *such that*

$$\Gamma_k \geq \underline{\Gamma}, \quad \Theta_k \geq \underline{\Theta}, \quad \text{and} \quad \Gamma_k\Theta_k \leq \Lambda, \tag{42}$$

we have

$$\begin{array}{lcl} \sum_{k=0}^{\infty}(k+1)\|G_\lambda x^k\|^2 & < & +\infty \ \ \text{almost surely}, \\ \sum_{k=0}^{\infty}(k+1)^2\|\widetilde{F}^k - Fx^k\|^2 & < & +\infty \ \ \text{almost surely}, \\ \sum_{k=0}^{\infty}(k+1)\|x^{k+1} - x^k\|^2 & < & +\infty \ \ \text{almost surely}. \end{array} \tag{43}$$

The following almost sure limits also hold (showing $o\big(1/k^2\big)$ *almost sure convergence rates):*

$$\begin{array}{ll} \lim_{k\to\infty} k^2\|G_\lambda x^k\|^2 & = 0, \\ \lim_{k\to\infty} k^2\|x^{k+1} - x^k\|^2 & = 0. \end{array} \tag{44}$$

Moreover, both $\{x^k\}$ *and* $\{z^k\}$ *almost surely converge to a* $\mathrm{zer}(\Phi)$*-valued random variable* $x^\star$ *as a solution of* (GE).

The second condition of (42) comes from the second condition of (34). As we will see from Subsections 4.4 and 4.5, there exists $\underline{\Theta} > 0$ such that $\Theta_k \geq \underline{\Theta}$ and we choose $\Gamma_k = \Gamma > 0$ for all $k \geq 0$ such that $\Gamma\Theta_k \leq \Lambda$. Thus, the condition (42) automatically holds.

Remark 15 (Sufficient condition for (38)**)** *Since* $t_k := \mu(k+r)$, *from* (38), *we have*

$$B_\infty = \frac{\mu^2\Lambda}{\beta}\sum_{k=0}^{\infty}\frac{(k+r-1)(k+r-2)\sigma_k^2}{\Theta_k}.$$

If $\Theta_k := \Theta > 0$ *is fixed for all* $k \geq 0$, *then a sufficient condition for* $B_\infty < +\infty$ *is either* $\sigma_k = 0$ *or* $\sigma_k^2 = \mathcal{O}\big(\frac{\sigma^2}{k^{3+\omega}}\big)$ *for any* $\omega > 0$ *and* σ^2 *given in Assumption 1.1*(ii). *If we use either SVRG or SARAH estimators to construct* $\widetilde{F}^k$, *then we need to choose an increasing mega-batch size* n_k *for evaluating* $\bar{F}\tilde{x}^k$ *or* $\bar{F}x^k$ *such that* $n_k := \mathcal{O}\big(k^{3+\omega}\big)$ *to guarantee* $B_\infty < +\infty$.

Remark 16 *In this paper, we do not consider the strong monotonicity of* Φ *in* (GE). *This case was studied in, e.g., Tran-Dinh (2024b) when* $T = 0$ *using a different class of unbiased variance-reduced estimators. We believe that our methods can be customized to handle the strong monotonicity of* Φ *and can achieve a linear convergence rate as in Tran-Dinh (2024b).*

Remark 17 *Our analysis above relies on the main inequality* (10) *for* $x = x^{k+1}$ *and* $y = x^k$, *and for* $x = x^k$ *and* $y = x^\star$. *We have proven that both* $\{\|x^{k+1} - x^k\|^2\}$ *and* $\{\|x^k - x^\star\|^2\}$ *almost surely converge to zero (the former converges with a* $o\big(1/k^2\big)$ *rate). Consequently, when* k *is sufficiently large, both* $\|x^{k+1} - x^k\|$ *and* $\|x^k - x^\star\|$ *are almost surely sufficiently small. Thus, we just require* (10) *to holds locally, which can be guaranteed if* T *is locally* ρ*-co-hypomonotone. This weaker condition expands the potential applicability of our methods to locally* ρ*-co-hypomonotone operators* T *in* (GE).

4.4 Complexity of VFOSA$_+$ for specific estimators in the finite-sum setting

In this section, we apply Theorem 12 to concrete estimators described in Section 3 to obtain explicit complexity bound for three cases: L-SVRG, SAGA, and L-SARAH. The parameters and constants β, μ, ν, r, Λ, and Ψ_0 used in what follows are given in Theorem 12. The proof of these results can be found in Appendices D.1.1, D.1.2, and D.1.3, respectively.

Corollary 18 (L-SVRG) *Suppose that Assumptions 1.1 and 1.2 hold for* (GE) *in the finite-sum setting* (F). *Let* $\{(x^k, y^k, z^k)\}$ *be generated by* (VFOSA$_+$) *using* t_k, λ, *and* β *as in Theorem 12. Let* $\widetilde{F}^k$ *be constructed by* (L-SVRG) *with* $\tilde{x}^0 = x^0$, $\bar{F}\tilde{x}^k := F\tilde{x}^k$, *and*

$$b_k := \lfloor c_b n^{2\omega} \rfloor \quad \text{and} \quad \mathbf{p}_k := \begin{cases} \frac{2}{c_p n^\omega} + \frac{4\mu}{\mu(k+r-1)-1} & \text{if } 0 \le k \le K_0 := \lfloor 4c_p n^\omega - r + 1 + \mu^{-1} \rfloor, \\ \frac{3}{c_p n^\omega} & \text{otherwise}, \end{cases}$$

where $c_p > 0$ *is a given constant,* $r > 5 + \frac{1}{\mu}$, $n^\omega \ge \frac{1}{c_p}\max\left\{\frac{2\mu(r-1)-2}{\mu(r-5)-1}, \frac{\mu(r-1)-1}{4\mu}\right\}$ *for a fixed* $\omega \in [0,1]$, *and* $c_b := \frac{5\beta c_p^2}{\mu\Lambda}$.

Then, for a given tolerance $\epsilon > 0$, *the expected total number* $\bar{\mathcal{T}}_K$ *of oracle calls* F_i *and* $J_{\lambda T}$ *evaluations to obtain* x^K *such that* $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$ *is at most*

$$\bar{\mathcal{T}}_K := \Big\lfloor n + 4n\big[2 + \ln(4c_p n^\omega)\big] + \tfrac{\sqrt{2}\Psi_0}{\mu\epsilon}\big(2c_b n^{2\omega} + \tfrac{3n^{1-\omega}}{c_p}\big)\Big\rfloor.$$

In particular, if we choose $\omega := \frac{1}{3}$, *then our oracle complexity is* $\bar{\mathcal{T}}_K = \mathcal{O}\big(n\ln(n) + \frac{n^{2/3}}{\epsilon}\big)$.

Note that when $\omega = \frac{1}{3}$, our mini-batch size b_k is $b_k = \mathcal{O}\big(n^{2/3}\big)$ and our probability $\mathbf{p}_k = \mathcal{O}\left(n^{-1/3}\right)$. The oracle complexity $\mathcal{O}\big(n\ln(n) + \frac{n^{2/3}}{\epsilon}\big)$ appears to be slightly worse than the $\mathcal{O}\big(n + \frac{n^{2/3}}{\epsilon}\big)$ bound obtained in Tran-Dinh (2024b) by a $\ln(n)$ factor. In our experiments, we often set $\mathbf{p}_k := \frac{1}{2n^{1/3}}$ and $b_k := \lfloor \frac{n^{2/3}}{2} \rfloor$, but we can appropriately adjust these parameters.

Corollary 19 (SAGA) *Suppose that Assumptions 1.1 and 1.2 hold for* (GE) *in the finite-sum setting* (F). *Let* $\{(x^k, y^k, z^k)\}$ *be generated by* (VFOSA$_+$) *using* t_k, λ, *and* β *from Theorem 12. Let* $\widetilde{F}^k$ *be constructed by* (SAGA) *with*

$$b_k := \begin{cases} 2c_b n^{2/3} + \frac{4\mu n}{\mu(k+r-1)-1} & \text{if } \ 0 \le k \le K_0 := \lfloor 4n^{1/3} + 1 + \mu^{-1} - r \rfloor, \\ 3c_b n^{2/3}, & \text{otherwise}, \end{cases}$$

where $c_b := \frac{5}{2}\sqrt{\frac{\beta}{\mu\Lambda}}$, $r > 5 + \frac{1}{\mu}$, *and* $n^{1/3} \ge \max\left\{\frac{2c_b[\mu(r-1)-1]}{\mu(r-5)-1}, \frac{\mu(r-1)-1}{4\mu}\right\}$.

Then, for a given tolerance $\epsilon > 0$, *the expected total number* $\bar{\mathcal{T}}_K$ *of oracle calls* F_i *and* $J_{\lambda T}$ *evaluations to obtain* x^K *such that* $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$ *is at most*

$$\bar{\mathcal{T}}_K := \Big\lfloor [8c_b + 4\ln(4n^{1/3})]n + \tfrac{3\sqrt{2}c_b\Psi_0 n^{2/3}}{\mu\epsilon}\Big\rfloor.$$

Again, the complexity of the SAGA variant stated in Corollary 19 is $\mathcal{O}\big(n\ln(n) + n^{2/3}\epsilon^{-1}\big)$, which is slightly worse than $\mathcal{O}\big(n + n^{2/3}\epsilon^{-1}\big)$ in Tran-Dinh (2024b). In our experiments, we often choose $b_k := \min\{n, \lfloor \frac{n^{2/3}}{2} \rfloor\}$, but we can appropriately adjust b_k.

Corollary 20 (L-SARAH) *Suppose that Assumptions 1.1 and 1.2 hold for* (GE) *in the finite-sum setting* (F). *Let* $\{(x^k, y^k, z^k)\}$ *be generated by* (VFOSA$_+$) *using* t_k, λ, *and* β *as in Theorem 12. Let* $\widetilde{F}^k$ *be constructed by* (L-SARAH) *with*

$$b_k := \lfloor c_b n^{\omega} \rfloor \quad \text{and} \quad \mathbf{p}_k := \begin{cases} \frac{1}{c_p n^{\omega}} + \frac{2\mu}{\mu(k+r-1)-1} & \text{if } 0 \le k \le K_0 := \lfloor 2c_p n^{\omega} - r + 1 + \mu^{-1} \rfloor, \\ \frac{2}{c_p n^{\omega}} & \text{otherwise}, \end{cases}$$

where $c_p > 0$ *is a given constant,* $r > 3 + \frac{1}{\mu}$, *and* $n^{\omega} \ge \frac{1}{c_p} \max\left\{\frac{\mu(r-1)-1}{\mu(r-3)-1}, \frac{\mu(r-1)-1}{2\mu}\right\}$ *for a fixed* $\omega \in [0,1]$, *and* $c_b := \frac{5\beta c_p}{\mu\Lambda}$.

Then, for a given tolerance $\epsilon > 0$, *the expected total number* $\bar{\mathcal{T}}_K$ *of oracle calls* F_i *and* $J_{\lambda T}$ *evaluations to obtain* x^K *such that* $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$ *is at most*

$$\bar{\mathcal{T}}_K := \Big\lfloor n + 2n\big[2 + \ln(2c_p n^{\omega})\big] + \tfrac{\sqrt{2}\Psi_0}{\mu\epsilon}\big(c_b n^{\omega} + \tfrac{2n^{1-\omega}}{c_p}\big) \Big\rfloor.$$

In particular, if we choose $\omega := \frac{1}{2}$, *then our oracle complexity is* $\bar{\mathcal{T}}_K := \mathcal{O}\big(n\ln(n^{1/2}) + \frac{\sqrt{n}}{\epsilon}\big)$.

In our experiments, we often choose $\mathbf{p}_k := \frac{1}{2n^{1/2}}$ and $b_k := \lfloor \frac{\sqrt{n}}{2} \rfloor$, but again, we can appropriately adjust these parameters. Note that we can also apply the HSGD estimator to the finite-sum setting (F) of (GE), but we still need a bounded variance assumption in order to estimate its oracle complexity. Nevertheless, we omit this result here.

Finally, we specify Theorem 13 and Theorem 14 for concrete estimators in Section 3. The following result is a direct consequence of Theorems 13 and 14 since $B_\infty = 0$.

Corollary 21 *For the finite-sum setting* (F) *of* (GE), *suppose that the L-SVRG, SAGA, and L-SARAH estimators are constructed as in Corollaries 18, 19, and 20, respectively. Then, the conditions* (34) *of Theorem 12 are fulfilled and* $B_\infty = 0$ *in* (38).

Consequently, the conclusions of both Theorem 13 and Theorem 14 are valid for (VFOSA$_+$) *using these three estimators.*

Let us discuss our results and compare them with existing works. When ϵ is small, the complexity of L-SARAH is better than that of L-SVRG and SAGA by a factor of $n^{1/6}$. The complexity of the L-SARAH variant is the same as the stochastic Halpern method in Cai et al. (2024). Nevertheless, our method is different from Cai et al. (2024) and we also achieve better $o\big(1/k^2\big)$ convergence rates, several summability results, almost sure convergence rates, and the almost sure convergence of iterates to a solution of (GE).

4.5 Complexity of VFOSA$_+$ for specific estimators in the expectation setting

Now, we derive the oracle complexity of (VFOSA$_+$) to solve (GE) in the expectation setting (E). We have three variants corresponding to the L-SVRG, L-SARAH, and HSGD estimators. The proof of these results are given in Appendicies D.2.1, D.2.2, and D.2.3.

Corollary 22 (L-SVRG) *Suppose that Assumptions 1.1 and 1.2 hold for* (GE) *in the expectation setting* (E). *Let* $\{(x^k, y^k, z^k)\}$ *be generated by* (VFOSA$_+$) *using* t_k, λ, *and* β *as in Theorem 12. Let* $\widetilde{F}^k$ *be constructed by* (L-SVRG) *with*

$$b_k = b := \big\lfloor \tfrac{c_b}{\epsilon^2} \big\rfloor, \quad n_k = n := \big\lfloor \tfrac{c_n}{\epsilon^3} \big\rfloor, \quad \text{and} \quad \mathbf{p}_k := 2\epsilon + \tfrac{4\mu}{\mu(k+r-1)-1},$$

where $r > 5 + \frac{1}{\mu}$, $\epsilon \in \big(0, \frac{\mu(r-5)-1}{2\mu(r-1)-2}\big]$, $c_b := \frac{10\beta}{\mu\Lambda}$, *and* $c_n := 12\sigma^2\max\{1, \frac{2\sqrt{2}\Psi_0}{\mu^2}\}$.

Then, for a given $\epsilon > 0$, *the expected total number* $\bar{\mathcal{T}}_K$ *of oracle calls* $\mathbf{F}(\cdot,\xi)$ *and* $J_{\lambda T}$ *evaluations to obtain* x^K *such that* $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$ *is at most* $\bar{\mathcal{T}}_K := \mathcal{O}\big(\epsilon^{-3} + \epsilon^{-3}\ln(\epsilon^{-1})\big)$.

Corollary 23 (L-SARAH) *Suppose that Assumptions 1.1 and 1.2 hold for* (GE) *in the expectation setting* (E). *Let* $\{(x^k, y^k, z^k)\}$ *be generated by* (VFOSA$_+$) *using* t_k, λ, *and* β *as in Theorem 12. Let* $\widetilde{F}^k$ *be constructed by* (L-SARAH) *with*

$$b_k = b := \big\lfloor \tfrac{c_b}{\epsilon} \big\rfloor, \quad n_k = n := \big\lfloor \tfrac{c_n}{\epsilon^3} \big\rfloor, \quad \text{and} \quad \mathbf{p}_k := \epsilon + \tfrac{2\mu}{\mu(k+r-1)-1},$$

where $r > 3 + \frac{1}{\mu}$, $\epsilon \in \big(0, \frac{\mu(r-3)-1}{\mu(r-1)-1}\big]$, $c_b := \frac{5\beta}{\mu\Lambda}$, *and* $c_n := 24\sigma^2\max\Big\{\frac{\sqrt{2}\Psi_0}{\mu^2}, \frac{1}{\mu\beta}\Big\}$.

Then, for a given tolerance $\epsilon > 0$, *the expected total number* $\bar{\mathcal{T}}_K$ *of oracle calls* $\mathbf{F}(\cdot,\xi)$ *and* $J_{\lambda T}$ *evaluations to obtain an approximate solution* x^K *such that* $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$ *is at most* $\bar{\mathcal{T}}_K := \mathcal{O}\big(\epsilon^{-2} + \epsilon^{-3} + \epsilon^{-3}\ln(\epsilon^{-1})\big)$.

Unlike the variance-reduced Halpern fixed-point method in Cai et al. (2022a), we choose a fixed mini-batch size b_k instead of varying it. It leads to a log-factor in our complexity.

Corollary 24 (HSGD) *Suppose that Assumptions 1.1 and 1.2 hold for* (GE) *in the expectation setting* (E). *Let* $\{(x^k, y^k, z^k)\}$ *be generated by* (VFOSA$_+$) *using the parameters* t_k, λ, *and* β *as in Theorem 12. Let* $\widetilde{F}^k$ *be constructed by* (HSGD) *with*

$$\begin{aligned}
&\tau_k := 1 - \sqrt{\tfrac{(1-\theta)t_{k-1}(t_{k-1}-1)}{t_k(t_k-1)}} \quad \text{for} \quad \theta := \epsilon,\\
&b_k = b := \big\lfloor \tfrac{c_b}{\epsilon} \big\rfloor, \quad \hat{b}_k = \hat{b} := \big\lfloor \tfrac{\hat{c}_b}{\epsilon^2} \big\rfloor \quad \text{and} \quad n_0 := \big\lfloor \tfrac{c_n}{\epsilon} \big\rfloor,
\end{aligned}$$

where $\epsilon \in (0, \frac{1}{2}]$ *is a given tolerance,* $r \ge 5 + \frac{1}{\mu}$, *and* c_b, $\hat{c}_b$, *and* c_n *are given constants.*

Then, for $\epsilon > 0$ *given above, the expected total number* $\bar{\mathcal{T}}_K$ *of oracle calls* $\mathbf{F}(\cdot,\xi)$ *and* $J_{\lambda T}$ *evaluations to obtain* x^K *such that* $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$ *is at most* $\bar{\mathcal{T}}_K := \mathcal{O}\big(\epsilon^{-3}\big)$.

Three variants: L-SVRG, L-SARAH, and HSGD, discussed in this subsection offer either $\widetilde{\mathcal{O}}(\epsilon^{-3})$ or $\mathcal{O}\big(\epsilon^{-3}\big)$ oracle complexity to achieve an ϵ-solution x^K. However, each variant has its own advantages and disadvantages. For instance, L-SVRG seems to have the worst complexity among three variants, but it is unbiased, which is often preferable in practice. HSGD has the best complexity of $\mathcal{O}\big(\epsilon^{-3}\big)$. The chosen mini-batch b in L-SARAH is $\mathcal{O}\big(\epsilon^{-1}\big)$ which is smaller than $\mathcal{O}\big(\epsilon^{-2}\big)$ in both L-SVRG and HSGD. Both L-SVRG and L-SARAH occasionally require mega-batches of the size $\mathcal{O}\big(\epsilon^{-3}\big)$ to evaluate the snapshot point with a probability $\mathbf{p}_k$, while HSGD does not need any mega-batch, except for the initial epoch. Note that our theoretical parameters given in this subsection aim at achieving the best theoretical complexity bounds. However, one can adjust the values of parameters r, β, b_k, $\hat{b}_k$, $\mathbf{p}_k$, and τ_k when implementing these methods, which may work better in practice.

5. Variance-Reduced Accelerated Backward-Forward Splitting Method

In this section, we explore the BFS mapping S_λ in (8) to develop an alternative algorithmic framework to solve (GE). This algorithm is less popular in optimization and can be viewed as a gradient-proximal method in contrast to the proximal-gradient method.

5.1 The algorithm and its implementation

Our method relies on approximating the BFS mapping $S_\lambda(\cdot)$ at u^k defined by (8) by a stochastic estimator $\widetilde{S}^k_\lambda$. This estimator is constructed as follows:

$$\widetilde{S}^k_\lambda := \widetilde{F}^k + \tfrac{1}{\lambda}(u^k - x^k), \tag{45}$$

where $x^k := J_{\lambda T}u^k$ and $\widetilde{F}^k$ is a variance-reduced estimator of Fx^k satisfying Definition 4. It is obvious to see that $\|\widetilde{S}^k_\lambda - S_\lambda u^k\| = \|\widetilde{F}^k - Fx^k\|$.

Now, we are ready to present the following algorithm, Algorithm 2, for solving (GE).

Algorithm 2 (**V**ariance-reduced **F**ast [BF] **O**perator **S**plitting **A**lgorithm (**VFOSA**$_-$))

Initialization: Take an initial point $x^0 \in \mathrm{dom}(\Phi)$.

  Choose μ, λ, and β as in Theorem 12. Set $\nu := \frac{\mu}{2}$ and $r := 2 + \frac{1}{\mu}$.

  Compute $u^0 := x^0 + \lambda\xi^0$ for any $\xi^0 \in Tx^0$ and set $s^0 := u^0$.

For $k := 0, \cdots, k_{\max}$ **do**

  Compute *the shadow iterate* $x^k := J_{\lambda T}u^k$.

  Construct an estimator $\widetilde{F}^k$ of Fx^k satisfying Definition 4.

  Update $t_k := \mu(k + r)$ and $\eta_k := \frac{2\beta(t_k - 1)}{t_k - \nu}$.

  Update the following steps:

$$\begin{cases} v^k &:= \frac{t_k - 1}{t_k}u^k + \frac{1}{t_k}s^k, \\ u^{k+1} &:= v^k - \eta_k\widetilde{S}^k_\lambda \equiv v^k - \frac{\eta_k}{\lambda}(u^k - x^k) - \eta_k\widetilde{F}^k, \\ s^{k+1} &:= s^k + \nu(u^{k+1} - v^k). \end{cases} \tag{VFOSA$_-$}$$

End For

Algorithm 2 also requires only one $\widetilde{F}^k$ evaluation and one $J_{\lambda T}$ evaluation per iteration. Hence, this method has the same per-iteration complexity as Algorithm 1 above.

5.2 Convergence analysis

Now, we apply the analysis in Section 4 to establish the convergence of Algorithm 2 in the following theorem, whose proof is given in Appendix E.1.

Theorem 25 *Suppose that Assumptions 1.1 and 1.2 hold for* (GE). *Let* $\{(x^k, u^k, v^k)\}$ *be generated by* (VFOSA$_-$) (*i.e., Algorithm 2*) *using an estimator* $\widetilde{F}^k$ *of* Fx^k *satisfying Definition 4. Under the same parameters* λ, ν, β, t_k *and* η_k *and the same conditions as in Theorem 12, let* S_λ *be defined by* (8) *and* $\xi^k := \frac{1}{\lambda}(u^k - x^k) \in Tx^k$. *Then, we have*

$$\mathbb{E}\big[\|Fx^K + \xi^K\|^2\big] \le \frac{2(\bar{\Psi}_0^2 + \bar{E}_0^2 + B_{K-1})}{\mu^2(K + r - 1)^2}, \tag{46}$$

where $\bar{\Psi}_0^2 := \mu^2r^2\mathbb{E}\big[\|Fx^0 + \xi^0\|^2\big] + \frac{2r-1}{4\beta^2(\mu r - 1)}\|u^0 - u^\star\|^2$, $\bar{E}_0^2 := \frac{\mu r^2(\mu\Gamma_0 + \beta)\sigma_0^2}{2\beta}$, *and* B_K *is given in Theorem 12.*

Moreover, we also have

$$\begin{aligned}\textstyle\sum_{k=0}^{K}\big[(2-3\mu)\mu(k+r)+6\mu^2\big]\mathbb{E}\big[\|Fx^k+\xi^k\|^2\big] &\leq 2\big(\bar{\Psi}_0^2+\bar{E}_0^2+B_K\big),\\ \textstyle\sum_{k=0}^{K}(k+r)(k+r-\mu^{-1})\mathbb{E}\big[\|\widetilde{F}^k-Fx^k\|^2\big] &\leq \tfrac{2(\bar{\Psi}_0^2+\bar{E}_0^2+B_K)}{\beta\mu}.\end{aligned} \tag{47}$$

Similarly, we can also state both $\mathcal{O}\left(1/k^2\right)$ and $o\big(1/k^2\big)$-convergence rates of (VFOSA$_-$), whose proof is very similar to the proof of Theorem 13, see Appendix E.2.

Theorem 26 *Under the same conditions and settings as in Theorems 13 and 25, if, in addition,* (38) *holds, then for* $\xi^k := \frac{1}{\lambda}(u^k - x^k) \in Tx^k$ *and for all* $k \geq 0$*, we have*

$$\mathbb{E}\big[\|Fx^k+\xi^k\|^2\big] \leq \frac{2(\bar{\Psi}_0^2+\bar{E}_0^2+B_\infty)}{\mu^2(k+r-1)^2}.$$

We have the following summability bounds:

$$\begin{aligned}\textstyle\sum_{k=0}^{\infty}(k+1)\mathbb{E}\big[\|Fx^k+\xi^k\|^2\big] &< +\infty,\\ \textstyle\sum_{k=0}^{\infty}(k+1)^2\mathbb{E}\big[\|\widetilde{F}^k-Fx^k\|^2\big] &< +\infty,\\ \textstyle\sum_{k=0}^{\infty}(k+1)\mathbb{E}\big[\|u^{k+1}-u^k\|^2\big] &< +\infty,\\ \textstyle\sum_{k=0}^{\infty}(k+1)\mathbb{E}\big[\|x^{k+1}-x^k\|^2\big] &< +\infty,\end{aligned}$$

where the last three bound require $0 < \beta < \frac{(2-\mu)\bar{\beta}}{2+\mu}$*. We also obtain the following limits:*

$$\begin{aligned}\textstyle\lim_{k\to\infty} k^2\mathbb{E}\big[\|u^{k+1}-u^k\|^2\big] &= 0,\\ \textstyle\lim_{k\to\infty} k^2\mathbb{E}\big[\|x^{k+1}-x^k\|^2\big] &= 0,\\ \textstyle\lim_{k\to\infty} k^2\mathbb{E}\big[\|Fx^k+\xi^k\|^2\big] &= 0.\end{aligned}$$

Theorem 27 *Under the same the conditions and settings as in Theorems 14 and 26, we have the following almost sure summability bounds:*

$$\begin{aligned}\textstyle\sum_{k=0}^{\infty}(k+1)\|Fx^k+\xi^k\|^2 &< +\infty,\\ \textstyle\sum_{k=0}^{\infty}(k+1)^2\|\widetilde{F}^k-Fx^k\|^2 &< +\infty,\\ \textstyle\sum_{k=0}^{\infty}(k+1)\|u^{k+1}-u^k\|^2 &< +\infty,\\ \textstyle\sum_{k=0}^{\infty}(k+1)\|x^{k+1}-x^k\|^2 &< +\infty.\end{aligned}$$

The following almost sure limits also hold (showing $o\big(1/k^2\big)$ *almost sure convergence rates):*

$$\lim_{k\to\infty} k^2\|x^{k+1}-x^k\|^2 = 0, \quad \textit{and} \quad \lim_{k\to\infty} k^2\|Fx^k+\xi^k\|^2 = 0.$$

Moreover, $\{u^k\}$ *almost surely converges to a* $\operatorname{zer}(S_\lambda)$*-valued random variable* $u^\star \in \operatorname{zer}(S_\lambda)$*. Consequently,* $\{x^k\}$ *also almost surely converges to* $x^\star := J_{\lambda T}u^\star \in \operatorname{zer}(\Phi)$*.*

If we specify (VFOSA$_-$) (or equivalently, Algorithm 2) to obtain a specific variant corresponding to each estimator in Section 3.2, then we still obtain the same oracle complexity as in (VFOSA$_+$). We omit these variants as they are similar to Subections 4.4 and 4.5.

Comparison between VFOSA$_+$ and VFOSA$_-$. Both VFOSA$_+$ and VFOSA$_-$ require the same assumptions to guarantee convergence. They also share the same convergence properties. However, $\{x^k\}$ is the primal sequence of VFOSA$_+$, and it directly converges to a solution of (GE) almost surely. The primal sequence of VFOSA$_-$ is $\{u^k\}$, which does not directly converge to a solution of (GE). Instead, the shadow sequence $\{x^k\}$ computed by $x^k := J_{\lambda T}u^k$ almost surely converges to a solution of (GE). As mentioned earlier, VFOSA$_+$ is more popular than VFOSA$_-$ in the literature. It covers the well-known proximal-gradient method in convex optimization as a special case.

6. Numerical Experiments

In this section, we present two numerical examples to validate our methods and compare the performance of different algorithms, including ours and recent methods from the literature. All the algorithms are implemented in Python and executed on a MacBook Pro with Apple M4 processor and 24Gb of memory.

6.1 Robust regularized logistic regression with ambiguous features

We apply our methods to solve the regularized logistic regression problem with ambiguous features studied in Tran-Dinh and Luo (2025) and compare them with the accelerated Halpern's fixed-point method using SARAH in Cai et al. (2024).

6.1.1 Mathematical model

We are given a dataset $\{(\hat{X}_i, y_i)\}_{i=1}^n$, where $\hat{X}_i$ is an i.i.d. sample of a feature vector in $\mathbb{R}^{p_1}$ and $y_i \in \{0, 1\}$ is the corresponding label of $\hat{X}_i$. We assume that $\hat{X}_i$ is ambiguous, i.e., it belongs to one of p_2 possible examples $\{X_{ij}\}_{j=1}^{p_2}$. Since we do not know $\hat{X}_i$ to evaluate the loss, we consider the worst-case loss $f_i(u) := \max_{1\le j\le p_2} \ell(\langle X_{ij}, u\rangle, y_i)$ computed from p_2 examples $\{X_{ij}\}_{j=1}^{p_2}$, where $\ell(\tau, s) := \log(1 + \exp(\tau)) - s\tau$ is the standard logistic loss.

Since $\max_{1\le j\le p_2} \ell_j(\cdot) = \max_{v\in\Delta_{p_2}} \sum_{j=1}^{p_2} v_j\ell_j(\cdot)$, where Δ_{p_2} is the standard simplex in $\mathbb{R}^{p_2}$, we can model this robustification of the regularized logistic regression problem into the following [non]convex-concave minimax formulation:

$$\min_{u\in\mathbb{R}^{p_1}} \Big\{\phi(u) := \max_{v\in\mathbb{R}^{p_2}} \Big\{\mathcal{H}(u, v) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p_2} v_j\ell(\langle X_{ij}, u\rangle, y_i) - \delta_{\Delta_{p_2}}(v)\Big\} + \bar{\lambda}R(u)\Big\}, \quad (48)$$

where $R(\cdot)$ is a given regularizer (possibly nonconvex) chosen later, $\bar{\lambda} > 0$ is a regularization parameter, and $\delta_{\Delta_{p_2}}$ is the indicator of Δ_{p_2} that handles the constraint $v \in \Delta_{p_2}$.

Let us denote by $x := [u; v]$ and

$$\begin{aligned} F_i x &:= \big[\textstyle\sum_{j=1}^{p_2} v_j\ell'(\langle X_{ij}, u\rangle, y_i)X_{ij};\ -\ell(\langle X_{i1}, u\rangle, y_i); \cdots; -\ell(\langle X_{ip_2}, u\rangle, y_i)\big], \\ Tx &:= [\bar{\lambda}\partial R(u); \partial\delta_{\Delta_{p_2}}(v)], \end{aligned}$$

where $\ell'(\tau, s) = \frac{\exp(\tau)}{1+\exp(\tau)} - s$. Then, we have $Fx = \frac{1}{n}\sum_{i=1}^n F_i x$ as given in (F). The optimality condition of (48) can be written as $0 \in Fx + Tx$, which is a special case of (GE).

6.1.2 Implementation details and input data

We implement both methods: VFOSA$_+$ and VFOSA$_-$ to solve (48). Each method consists of 4 different variants: L-SVRG, SAGA, L-SARAH, and Hybrid-SGD. They are abbreviated by `VFOSA`$_+$`-Svrg`, `VFOSA`$_+$`-Saga`, `VFOSA`$_+$`-Sarah`, `VFOSA`$_+$`-Hsgd`, `VFOSA`$_-$`-Svrg`, `VFOSA`$_-$`-Saga`, `VFOSA`$_-$`-Sarah`, and `VFOSA`$_-$`-Hsgd`, respectively. We consider two cases of (48) as follows.

- **The monotone case.** We choose $R(u) := \|u\|_1$ as an ℓ_1-norm regularizer, leading to a convex-concave saddle-point instance of (48). This corresponds to a monotone T in (GE). In this case, Assumption 1.1(iii) holds with $\rho = 0$.
- **The nonmonotone case**: We choose $R(u)$ to be the SCAD (smoothly clipped absolute deviation) regularizer from (Fan and Li, 2001), which promotes sparsity of u using a nonconvex regularizer. Hence, (48) is nonconvex-concave, leading to a nonmonotone T in (GE). In fact, T is locally ρ-co-hypomonotone with $\rho \le a - 1 = 2.7$. By Remark 17, we can still apply our methods to solve (48).

For comparison, we also implement the variance-reduced Halpern's fixed-point method from (Cai et al., 2024), which is abbreviated by `VrHalpern`. For further comparison with other methods such as extragradient, variance-reduced extragradient, and extra-anchor gradient methods (Yoon and Ryu, 2021), we refer to Subsection 6.2 and also (Cai et al., 2024).

Since it is difficult to exactly evaluate the co-coercivity constant L, we approximate it by $L := \frac{1}{4}\|X^T X\|$, which represents the smoothness constant of the logistic loss. Then, we choose $\lambda := \frac{1}{2L} < \frac{2}{L}$, and $\bar{\beta} := \frac{\lambda(4-L\lambda)}{4}$ as suggested by Lemma 3 and Theorem 12. From our theory and (Cai et al., 2024), we choose the parameters for each variant as follows.

- We choose $\mu := \frac{0.95\times 2}{3} < \frac{2}{3}$ and $r := 2 + \frac{1}{\mu}$ for all variants of our methods.
- For the L-SVRG variants, we choose $\mathbf{p}_k = \frac{1}{2n^{1/3}}$ and $b := \lfloor \frac{n^{2/3}}{2} \rfloor$, see Corollary 18.
- For the SAGA variants, we choose $b := \lfloor \frac{n^{2/3}}{2} \rfloor$, see Corollary 19.
- For the L-SARAH variants, we choose $\mathbf{p}_k := \frac{1}{2\sqrt{n}}$ and $b := \lfloor \frac{\sqrt{n}}{2} \rfloor$, see Corollary 20.
- For the Hybrid-SGD variants, we choose $\theta := \frac{1}{n}$ and $b := \lfloor \frac{\sqrt{n}}{2} \rfloor$, see Corollary 24.
- For `VrHalpern`, we choose $\mathbf{p}_k := \frac{1}{2\sqrt{n}}$ and $b := \lfloor \frac{\sqrt{n}}{2} \rfloor$, see (Cai et al., 2024).

We report the relative norm $\|G_\lambda x^k\|/\|G_\lambda x^0\|$ against the number of epochs as our main criterion in each experiment. We choose the initial point $x^0 := 0.25 \times \mathrm{randn}(p)$ in all methods, and run each algorithm for $N_e := 200$ epochs. Note that in the following experiments, we do not implement any tuning strategy for our parameters.

We use 4 different real datasets from `LIBSVM` (Chang and Lin, 2011): `gisette` (5,000 features and 6,000 samples), `w8a` (300 features and 49,749 samples), `a9a` (123 features and 32,561 samples), and `mnist` (784 features and 60,000 samples). Here, the first dataset has a large number of features but a small number of samples, while the other ones have a small number of features but a large number of samples. Since `mnist` is designed for multi-class classification, we convert it to a binary format by grouping the even digits (i.e., $\{0, 2, 4, 6, 8\}$) into one class and the odd digits (i.e., $\{1, 3, 5, 7, 9\}$) into the other. As usual, we normalize the feature vector $\hat{X}_i$ such that each sample has unit norm, and add a column of all ones to address the bias term.

To generate ambiguous features, we apply the following procedure. First, we choose the number of ambiguous features to be $p_2 = 10$ (10 groups). Next, for each sample i, we take the nominal feature vector $\hat{X}_i$ and add a random noise generated from a normal distribution

of zero mean and variance $\sigma^2 = 0.05^2$. We also choose the regularization parameter $\bar{\lambda}$ to be $\bar{\lambda} := 5 \times 10^{-3}$. This $\bar{\lambda}$ lead to a reasonable sparsity pattern of an approximate solution u^k to $u^\star$ of (48), though it does not produce a highly accurate residual norm.

6.1.3 Numerical experiments

We carry out different experiments to test our VFOSA$_+$ and VFOSA$_-$ for both the monotone and nonmonotone T using four estimators: L-SVRG, SAGA, L-SARAH, and HSGD.

(a) **Comparing 4 variants of VFOSA$_+$ on monotone problems.** First, we run four variants: L-SVRG, SAGA, L-SARAH, and HSGD on two real datasets: `w8a` and `gisette` to solve a convex-concave minimax instance of (48) with the ℓ_1-regularizer $R(u) = \|u\|_1$ (i.e., T is monotone). The results of this experiment are revealed in Figure 1.

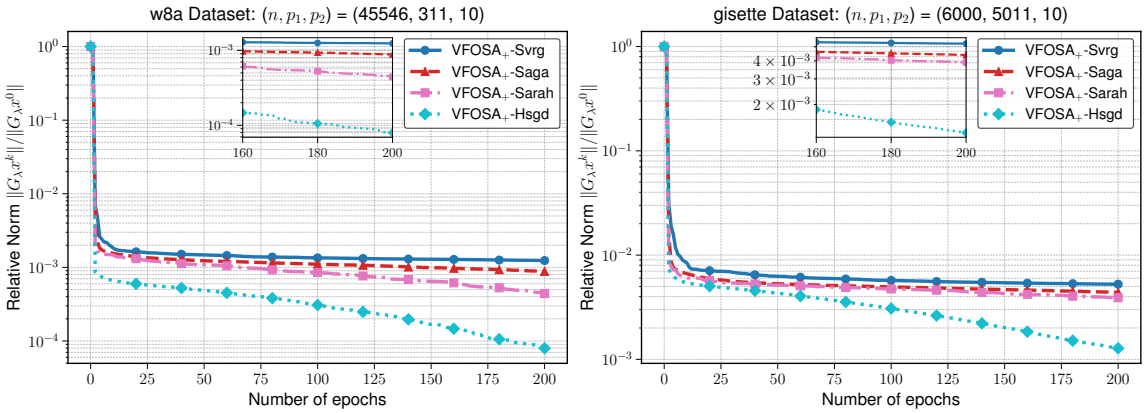

Figure 1: The performance of 4 variants of VFOSA$_+$: L-SVRG, SAGA, L-SARAH, and HSGD for solving (48) with the ℓ_1-regularizer on two datasets: `w8a` and `gisette`.

One can see from Figure 1 that both `VFOSA+-Svrg` and `VFOSA+-Saga` achieve similar performance, with `VFOSA+-Saga` slightly outperforming `VFOSA+-Svrg`, despite having the same order of oracle complexity. `VFOSA+-Sarah` performs better than both `VFOSA+-Svrg` and `VFOSA+-Saga`, while `VFOSA+-Hsgd` appears to outperform all its competitors. Although `VFOSA+-Svrg` and `VFOSA+-Sarah` occasionally compute full batches of F, `VFOSA+-Saga` and `VFOSA+-Hsgd` require no full-batch evaluations except for the first epoch. However, `VFOSA+-Saga` must store all component mappings $F_i x^k$. This experiment confirms that biased estimators such as L-SARAH and HSGD outperform unbiased ones like L-SVRG and SAGA, aligning with our theoretical findings as well as known results in optimization.

Similarly, we also compare these four algorithmic variants on the other two datasets: `mnist` and `a9a`. The results are reported in Figure 2.

Again, we observe a similar performance as in the first experiment. Both `VFOSA+-Sarah` and `VFOSA+-Hsgd` still outperform `VFOSA+-Svrg` and `VFOSA+-Saga`. Overall, `VFOSA+-Hsg` is still the best in this experiment.

(b) **Comparing 4 variants of VFOSA$_+$ on nonmonotone problems.** Next, we run these four algorithmic variants on four datasets above to solve a nonconvex-concave minimax instance of (48) by choosing the SCAD regularizer (i.e., nonmonotone T). The results are shown in Figure 3 and Figure 4, respectively for each pair of datasets.

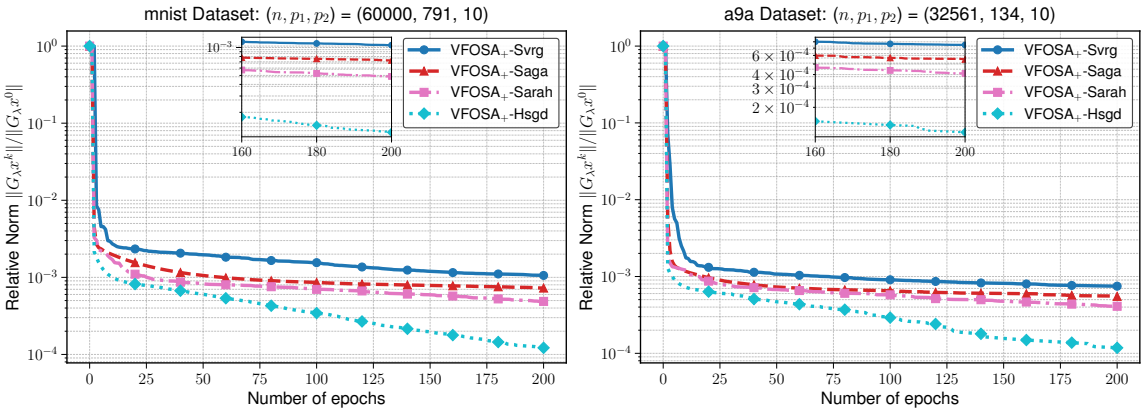

Figure 2: The performance of 4 variants of VFOSA$_+$: L-SVRG, SAGA, L-SARAH, and HSGD for solving (48) using the ℓ_1-regularizer on two datasets: `mnist` and `a9a`.

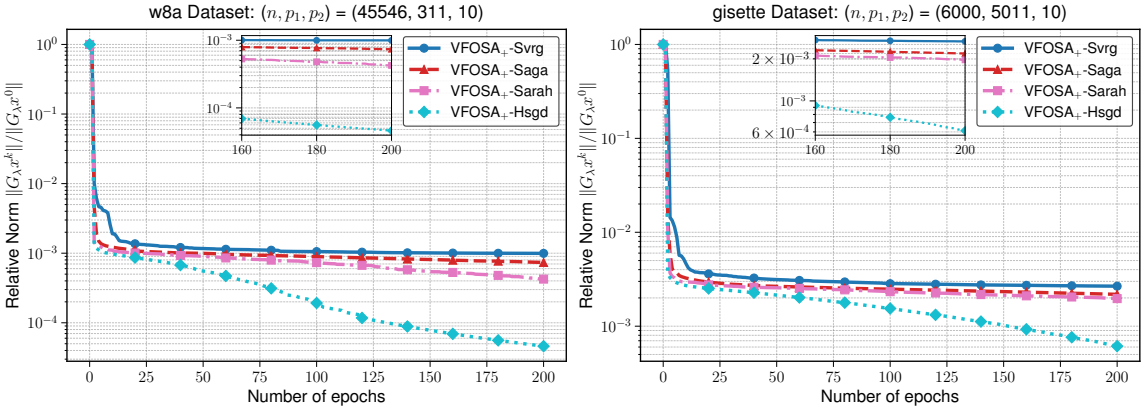

Figure 3: The performance of 4 variants of VFOSA$_+$: SVRG, SAGA, SARAH, and HSGD for solving (48) using the SCAD regularizer on two datasets: `w8a` and `gisette`.

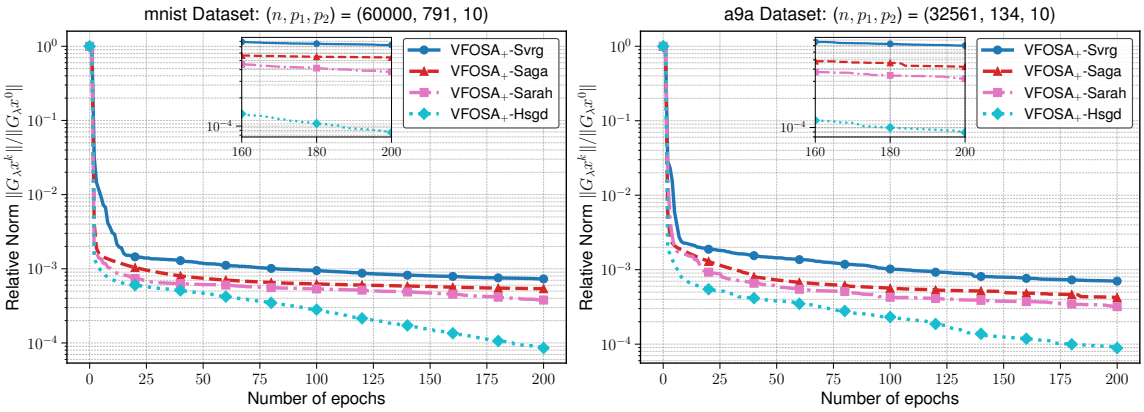

Figure 4: The performance of 4 variants of VFOSA$_+$: L-SVRG, SAGA, L-SARAH, and HSGD for solving (48) using the SCAD regularizer on the `mnist` and `a9a` datasets.

We can see from both Figures 3 and 4 that our algorithms still work well with the SCAD regularizer, and show a similar performance as our first two experiments. Nevertheless, the solutions we obtain have a better sparsity pattern compared to the ℓ_1-norm regularizer.

(c) **Comparing 4 variants of VFOSA$_-$.** Alternatively, we conduct a similar type of experiments for our VFOSA$_-$ method. This time we only compare four variants of VFOSA$_-$ on two datasets: `mnist` and `gisette` for both the ℓ_1-norm regularizer (monotone case) and the SCAD regularizer (nonmonotone case). The results of this experiment are presented in Figures 5 and 6, respectively.

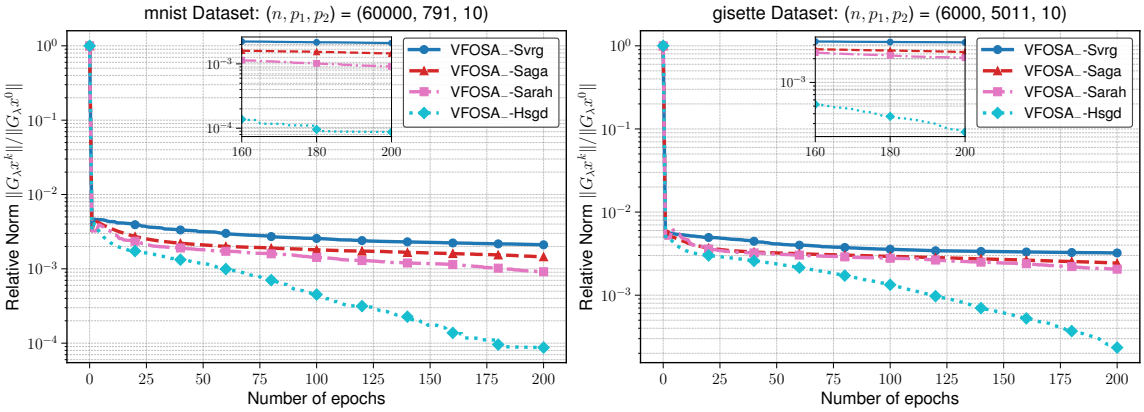

Figure 5: The performance of 4 variants of VFOSA$_-$: L-SVRG, SAGA, L-SARAH, and HSGD for solving (48) with the ℓ_1-norm regularizer on `mnist` and `gisette`.

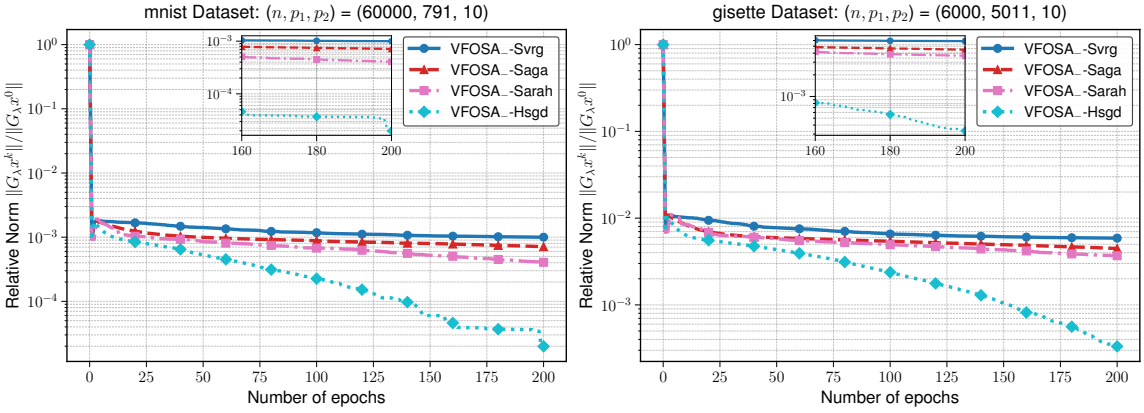

Figure 6: The performance of 4 variants of VFOSA$_-$: L-SVRG, SAGA, L-SARAH, and HSGD for solving (48) with the SCAD regularizer on `mnist` and `gisette`.

We still see that our L-SARAH and HSGD work better than L-SVRG and SAGA. They quickly reach the 10^{-2} to 10^{-3} accuracies after a few epochs, but then make slow progress later. HSGD still outperforms its competitors, especially on the `gisette` and `a9a` datasets.

(d) **Comparing VFOSA$_+$, VFOSA$_-$, and `VrHalpern`.** Finally, we compare our methods: VFOSA$_+$ and VFOSA$_-$ and `VrHalpern` in (Cai et al., 2024). For each of our methods, we choose three variants: L-SVRG, L-SARAH, and HSGD as they do not need to store F_i as in SAGA. We run these algorithms on two datasets: `mnist` and `gisette` with the same

setting as in the previous experiments for both the monotone and nonmonotone cases. For `VrHalpern`, we choose $\lambda_k := \frac{2}{k+4}$ and $\eta := \frac{1}{2L}$. This η is consistent to our learning rates, but twice larger than the suggested value $\eta = \frac{1}{4L}$ in (Cai et al., 2024).

We run this experiment for $N_e = 200$ epochs as before using both the ℓ_1-norm and SCAD regularizers. Figures 7 and 8 reveal the results of these methods for each case, respectively.

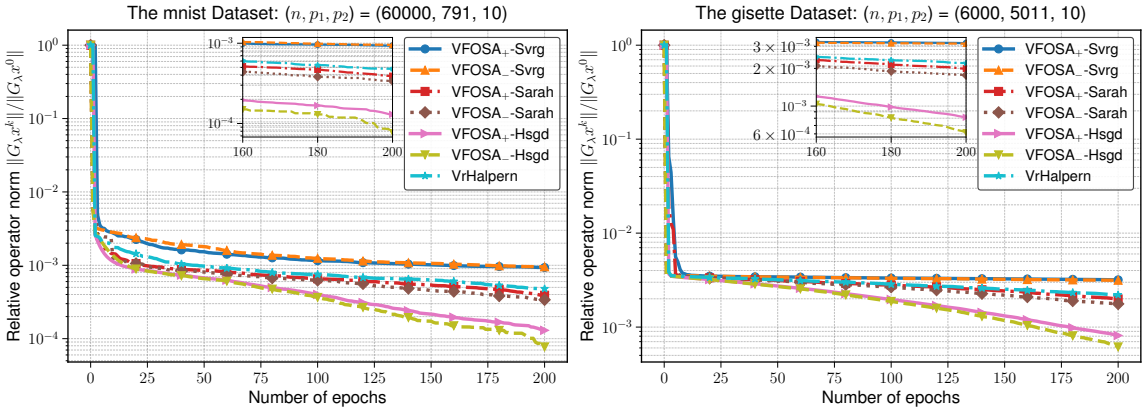

Figure 7: The performance of 7 algorithms: 3 variants of each VFOSA$_+$ and VFOSA$_-$, and `VrHalpern` using the ℓ_1-regularizer on two datasets: `mnist` and `gisette`.

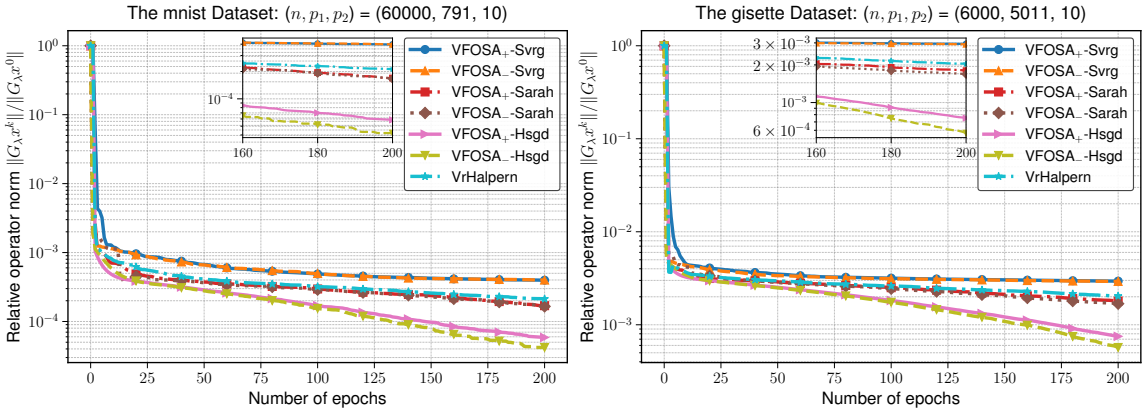

Figure 8: The performance of 7 algorithms: 3 variants of each VFOSA$_+$ and VFOSA$_-$, and `VrHalpern` using the SCAD regularizer on 2 datasets: `mnist` and `gisette`.

As observed in both Figures 7 and 8, our L-SARAH and HSGD variants perform better than L-SVRG and `VrHalpern` on the `mnist` and `gisette` datasets. They also outperform L-SVRG and `VrHalpern` in both the monotone and nonmonotone cases. Our `VFOSA_-Hsgd` appears to perform slightly better than `VFOSA_+-Hsgd`. A key reason for the superior performance of the HSGD variants is that they avoid full-batch computations except for the first epoch compared to the L-SARAH variant, reducing the number of F_i evaluations. `VrHalpern` performs better than our L-SVRG variants, which aligns with the theoretical complexity results in both methods.

6.2 Policeman vs. Burglar problem: Comparison between different methods

(a) **Mathematical model.** Following Nemirovski (2013), we consider the Policeman vs. Burglar problem as follows. There are p_1 houses in a city, where the i-th house has wealth $w_i \in \mathbb{R}_+$. Every evening, the Burglar chooses a house i to attack, and the Policeman chooses his post near a house j for all $1 \le i, j \le p_1$. After the burglary begins, the Policeman becomes aware of where it is happening, and his probability of catching the Burglar is $\mathbf{p}_c := \exp\{-\theta \cdot \mathrm{dist}(i,j)\}$, where $\mathrm{dist}(i,j)$ is the distance between houses i and j. On the other hand, the Burglar seeks to maximize his expected profit $\mathbf{L}_{ij} = w_i\big(1 - \exp\{-\theta \cdot \mathrm{dist}(i,j)\}\big)$, while the Policeman's interest is completely opposite.

Let $\mathbf{L}$ be a $p_1 \times p_1$ symmetric matrix such that $\mathbf{L}_{ij} := w_i(1 - \exp\{-\theta \cdot \mathrm{dist}(i,j)\})$ for $1 \le i, j \le p_1$, and let $\mathcal{U} = \mathcal{V} := \Delta_{p_1}$ be the standard simplex in $\mathbb{R}^{p_1}$. Then, the above Policeman vs. Burglar problem can be formulated into the following two-person game:

$$\min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} \big\{\mathcal{H}(u,v) := \langle \mathbf{L}u, v\rangle\big\}, \tag{49}$$

where u and v represent mixed strategies of the Policeman and the Burglar, respectively.

The optimality condition of (49) is $0 \in Fx + Tx$, which is a special case of (GE), where $x := [u, v]$, $Fx := [\mathbf{L}^\top v; -\mathbf{L}u]$, and $Tx := [\partial\delta_{\mathcal{U}}(u), \partial\delta_{\mathcal{V}}(v)]$ with $\delta_{\mathcal{X}}$ being the indicator of $\mathcal{X}$. Clearly, F is skew-symmetric and thus monotone, but it is not average co-coercive. To make F average co-coercive, we add a small regularizer so that $Fx = [\epsilon u + \mathbf{L}^\top v; \epsilon v - \mathbf{L}u]$ for $\epsilon = 10^{-8}$. This modification does not significantly interfere with (49).

(b) **Generating input data.** First, we choose $\mathrm{dist}(i,j) := |i - j|$ and set $\theta := 0.8$ that reflects a reasonable probability $\mathbf{p}_c$ of catching the Burglar. Next, we generate a vector $\hat{w} \in \mathbb{R}^{p_1}_+$ randomly from a standard normal distribution, followed by taking the absolute value to ensure nonnegativity. We call this vector the nominal wealth. Then, the wealth vector w is generated by $w := |\hat{w} + \sigma \cdot \texttt{randn}(q)|$ as a nonnegative random vector, where $\sigma^2 := 0.05$ is the variance of the noise. Now, assume that $\mathbf{L} := \frac{1}{n}\sum_{s=1}^n \mathbf{L}_s$ is the mean of n samples $\mathbf{L}_s$ generated from the samples w_s of w for $s = 1, \cdots, n$. Using this procedure, we generate two sets of problems corresponding to the two experiments as follows.

- *Experiment 1.* Choose $p_1 := 100$ houses (on a 10×10 grid) and $n = 1000$ samples, and generate 10 problem instances of size $p = 2p_1 = 200$.
- *Experiment 2.* Choose $p_1 := 225$ houses (on a 15×15 grid) and $n = 2000$ samples, and also generate 10 problem instances of size $p = 2p_1 = 450$.

(c) **Our algorithms and their competitors.** We select the following five competitors.

- The optimistic gradient method, e.g., in (Daskalakis et al., 2018), abbreviated by `OG`. It is a non-accelerated deterministic variant of Popov's past-extragradient method.
- The fast Krasnosel'kiĭ-Mann (KM) method in Bot and Nguyen (2022); Tran-Dinh (2024a), called `FKM`. This is a Nesterov's accelerated variant of the KM scheme.
- The variance-reduced forward-reflected-backward splitting (FRBS) algorithm in Alacaoglu et al. (2021), abbreviated by `VrFRBS`.
- The variance-reduced extragradient (EG) algorithm in Alacaoglu and Malitsky (2022), abbreviated by `VrEG`. This is a non-accelerated variance-reduced EG method.
- The variance-reduced Halpern fixed-point method in Cai et al. (2024), called `VrHalpern`.

Since the stochastic competitors either use L-SVRG or L-SARAH, we implement two variants of our VFOSA$_+$: `VFOSA`$_+$`-Svrg` and `VFOSA`$_+$`-Sarah` and compare them with the above

competitors. We run each experiment on 10 problem instances and report the mean of the relative FBS residual norm $\|G_\lambda x^k\|/\|G_\lambda x^0\|$ against the number of epochs for 200 epochs.

(d) **Parameter selection.** We test all the algorithms using the recommended parameters from their theory. More specifically, for all the L-SVRG variants, we choose the probability for snapshot points $\tilde{x}^k$ as $\mathbf{p}_k = \frac{1}{2n^{1/3}}$ and the mini-batch size $b_k := \lfloor \frac{n^{2/3}}{2} \rfloor$, while for all the L-SARAH variants, we choose $\mathbf{p}_k = \frac{1}{2\sqrt{n}}$ and $b_k := \lfloor \frac{\sqrt{n}}{2} \rfloor$. We also choose $x^0 := \frac{2}{p} \cdot \text{ones}(p)$ as the initial point in all methods so that it is feasible to (49). The learning rate of both `OG` and `FKM` is $\eta = \frac{1}{L}$ as suggested by their theory. The learning rate of `VrFRBS` is $\eta = 0.99 \cdot \frac{1-\sqrt{1-\mathbf{p}_k}}{2L}$ as recommended in Alacaoglu et al. (2021). The learning rate of `VrEG` is $\eta = 0.99 \cdot \frac{\sqrt{1-\alpha}}{L}$ for $\alpha := 1 - \mathbf{p}_k$ as shown in Alacaoglu and Malitsky (2022). Note that the learning rate of both `VrFRBS` and `VrEG` was derived for the single sample case, while we use it here for the mini-batch case. However, since $\mathbf{p}_k$ is larger than the theoretical value $\frac{1}{n}$ or $\frac{2}{n}$, this learning rate is larger than the one in their paper. The learning rate of `VrHalpern` is $\eta := \frac{1}{4L}$ as in Cai et al. (2024). For our VFOSA$_+$ methods, since $\rho = 0$, we choose $\mu := 0.95 \cdot \frac{2}{3}$, $r := 2 + \frac{1}{\mu}$, $\lambda := \frac{1}{L}$, $\bar{\beta} := \frac{\lambda(4-L\lambda)}{4}$, and $\beta := \frac{(2-\mu)\bar{\beta}}{2+\mu}$ as suggested by our theory in Theorem 12.

(e) **Numerical results.** The performance of all the algorithms is reported in Figure 9.

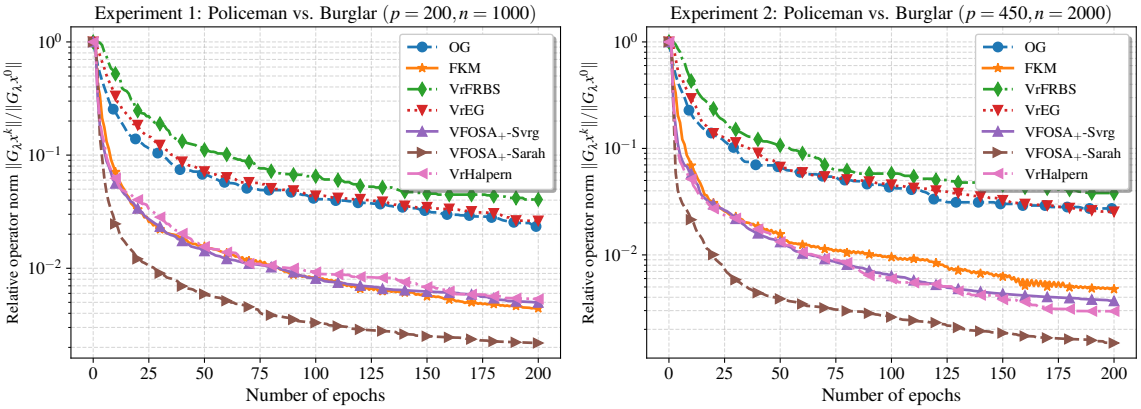

Figure 9: The performance of 2 variants of VFOSA$_+$ and 5 competitors for solving (49) using theoretical parameters. The average of 10 problems in each experiment.

As shown in Figure 9, the three accelerated methods consistently outperform their non-accelerated counterparts, including both deterministic and variance-reduced variants. The three non-accelerated schemes: `OG`, `VrFRBS`, and `VrEG`, exhibit similar performance across both experiments when using their respective theoretical parameters. Among the accelerated methods, the deterministic algorithm `FKM` still performs well and is comparable to our `VFOSA`$_+$`-Svrg` and `VrHalpern`. However, our `VFOSA`$_+$`-Sarah` achieves the best performance, attaining the lowest relative residual norm among all the methods.

Finally, we reduce both $\mathbf{p}_k$ and b_k by half to obtain a smaller probability and mini-batch size, respectively. We then rerun both experiments to evaluate how these parameters influence the performance of the stochastic methods. The results are presented in Figure 10.

As shown in Figure 10, the variance-reduced accelerated methods show an improved performance compared to the previous run. They also outperform the deterministic `FKM`

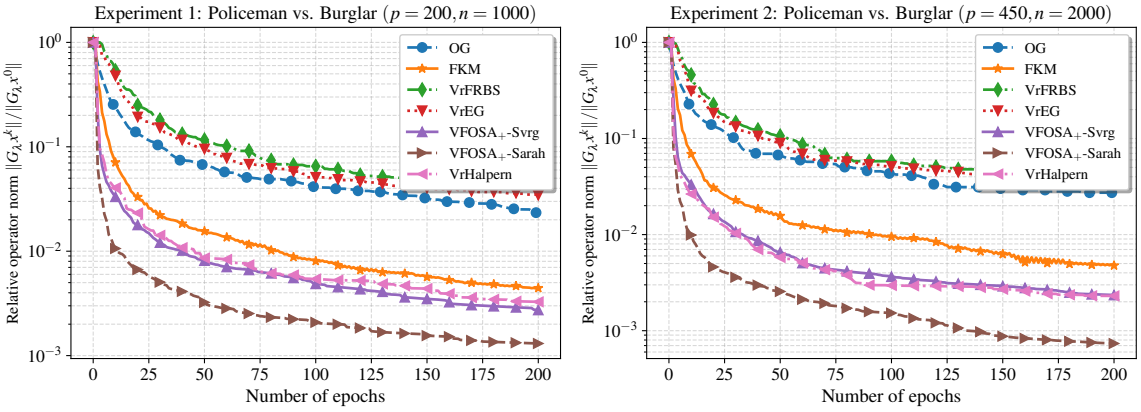

Figure 10: The performance of 2 variants of VFOSA$_+$ and 5 competitors for solving (49) using smaller $\mathbf{p}_k$ and b_k. The average of 10 problems in each experiment.

method from the earlier experiment. This is because the smaller values of $\mathbf{p}_k$ and b_k increase the number of iterations within the same number of epochs, enhancing performance.

7. Conclusions

We have developed two fast operator splitting frameworks with variance reduction to solve a class of generalized equations in both finite-sum and expectation settings, covering certain nonmonotone problems. Our methods exploit both the forward-backward and backward-forward splitting schemes and support a wide range of variance-reduced estimators, covering both unbiased and biased instances. We have established convergence rates of order $\mathcal{O}(1/k^2)$ and $o(1/k^2)$ in expectation for our methods. Then, we have also proved almost sure $o(1/k^2)$ convergence rates along with almost sure convergence of the iterates to a solution of our problem. Our frameworks comprise popular estimators such as L-SVRG, SAGA, L-SARAH, and Hybrid-SGD, and we have derived oracle complexity results that match or closely approach the best-known ones for optimization methods in the literature. Several interesting questions remain open. For example, can our approach be extended to extragradient methods and their variants to weaken the co-coercivity of F? Can adaptive schemes be developed to remove the need for estimating the co-coercivity constant L and the co-hypomonotonicity modulus ρ? We plan to explore these topics in our future work.

Acknowledgements. This work is partially supported by the National Science Foundation (NSF), grant no. NSF-RTG DMS-2134107 and the Office of Naval Research (ONR), grant No. N00014-23-1-2588 (2023-2026). The author gratefully acknowledges Mr. Nghia Nguyen-Trung for his careful proofreading of this work.

Appendix A. Technical Lemmas and Proof of Lemma 3

This appendix recalls necessary technical results and gives the full proof of Lemma 3.

A.1 Technical lemmas

We need the following technical results for our convergence analysis in the sequel.

Lemma 28 (Bauschke and Combettes (2017), Lemma 5.31) *Let $\{\alpha_k\}$, $\{\zeta_k\}$, $\{\gamma_k\}$, and $\{\varepsilon_k\}$ be nonnegative sequences such that $\sum_{k=0}^{\infty}\gamma_k < +\infty$ and $\sum_{k=0}^{\infty}\varepsilon_k < +\infty$. In addition, for all $k \geq 0$, we assume that*

$$\alpha_{k+1} \leq (1+\gamma_k)\alpha_k - \zeta_k + \varepsilon_k. \tag{50}$$

Then, we conclude that $\lim_{k\to\infty}\alpha_k$ exists and $\sum_{k=0}^{\infty}\zeta_k < +\infty$.

Lemma 29 *Given a nonnegative sequence $\{\alpha_k\}$ and $\omega \geq 0$ such that $\lim_{k\to\infty} k^{\omega+1}\alpha^k$ exists and $\sum_{k=0}^{\infty} k^{\omega}\alpha_k < +\infty$. Then, we conclude that $\lim_{k\to\infty} k^{\omega+1}\alpha^k = 0$.*

Proof Since $\alpha_k \geq 0$, suppose by contradiction that $\lim_{k\to\infty} k^{\omega+1}\alpha^k = \alpha > 0$. For any $0 < \epsilon < \alpha$, there exists k_0 sufficiently large such that $k^{\omega+1}\alpha_k \geq \alpha - \epsilon > 0$ for all $k \geq k_0$. Hence, we get $k^{\omega}\alpha_k \geq \frac{\alpha-\epsilon}{k}$. However, since $\sum_{k=0}^{\infty} k^{\omega}\alpha_k < +\infty$, the last relation leads to

$$+\infty < \textstyle\sum_{k=k_0}^{\infty}\frac{\alpha-\epsilon}{k} \leq \sum_{k=k_0}^{\infty} k^{\omega}\alpha_k < +\infty.$$

This relation shows a contradiction. Thus, we conclude that $\lim_{k\to\infty} k^{\omega+1}\alpha^k = \alpha = 0$. ■

We also need the well-known Robbins-Siegmund supermartingale theorem (Robbins and Siegmund, 1971), which we state it here as a technical lemma.

Lemma 30 *Let $\{U_k\}$, $\{\gamma_k\}$, $\{V_k\}$ and $\{E_k\}$ be sequences of nonnegative integrable random variables on some arbitrary probability space and adapted to the filtration $\{\mathcal{F}_k\}_{k\geq 0}$ with $\sum_{k=0}^{\infty}\gamma_k < +\infty$ and $\sum_{k=0}^{\infty} E_k < +\infty$ almost surely, and*

$$\mathbb{E}\big[U_{k+1} \mid \mathcal{F}_k\big] \leq (1+\gamma_k)U_k - V_k + E_k, \tag{51}$$

almost surely for all $k \geq 0$. Then, $\{U_k\}$ almost surely converges to a random variable and $\sum_{k=0}^{\infty} V_k < +\infty$ almost surely.

The following lemma is Proposition 4.1 of Davis (2022). It was proven for a demiclosed mapping G, but we recall it here for the case G is continuous in a finite-dimensional space.

Lemma 31 (Davis (2022), Proposition 4.1) *Suppose that $G : \mathbb{R}^p \to \mathbb{R}^p$ is continuous and $\operatorname{zer}(G) \neq \emptyset$. Let $\{x^k\}$ be a sequence of random vectors such that for all $x^\star \in \operatorname{zer}(G)$, the sequence $\{\|x^k - x^\star\|\}$ almost surely converges to a $[0,\infty)$-valued random variable. In addition, assume that $\{\|Gx^k\|\}$ also almost surely converges to zero. Then, $\{x^k\}$ almost surely converges to a $\operatorname{zer}(G)$-valued random variable.*

A.2 The proof of Lemma 3: Equivalent reformulations of (GE)

Proof Let $A_{\lambda T}x := \frac{1}{\lambda}(x - J_{\lambda T}x)$ be the Moreau-Yosida approximation of λT. First, we show that $A_{\lambda T}$ is $(\lambda-\rho)$-co-coercive, provided that $\lambda > \rho$. The co-coercivity of $A_{\lambda T}$ was proven, e.g., in Attouch et al. (2018), but we give a short proof here for completeness.

Indeed, for any x and y, we denote by $u := J_{\lambda T}x$ and $v := J_{\lambda T}y$. Then, we have $A_{\lambda T}x = \frac{1}{\lambda}(x-u) \in Tu$ and $A_{\lambda T}y = \frac{1}{\lambda}(y-v) \in Tv$. Since T is ρ-co-hypomonotone,

we have $\langle A_{\lambda T}x - A_{\lambda T}y, u - v\rangle \geq -\rho\|A_{\lambda T}x - A_{\lambda T}y\|^2$. Substituting $u = x - \lambda A_{\lambda T}x$ and $v = y - \lambda A_{\lambda T}y$ into this inequality, and rearranging the result, we get $\langle A_{\lambda T}x - A_{\lambda T}y, x - y\rangle \geq (\lambda - \rho)\|A_{\lambda T}x - A_{\lambda T}y\|^2$, which proves that $A_{\lambda T}$ is $(\lambda - \rho)$-co-coercive, provided that $\lambda > \rho$.

(i) To prove (10), from (6) we have $A_{\lambda T}(x - \lambda Fx) = \frac{1}{\lambda}(x - \lambda Fx - J_{\lambda T}(x - \lambda Fx)) = G_\lambda x - Fx$ and $A_\lambda(y - \lambda Fy) = G_\lambda y - Fy$. By the $(\lambda - \rho)$-co-coercivity of $A_{\lambda T}$, we have

$$\langle G_\lambda x - G_\lambda y - (Fx - Fy), x - y - \lambda(Fx - Fy)\rangle \geq (\lambda - \rho)\|G_\lambda x - G_\lambda y - (Fx - Fy)\|^2.$$

Expanding this inequality and rearranging the result, we get

$$\begin{aligned}\langle G_\lambda x - G_\lambda y, x - y\rangle &\geq (\lambda - \rho)\|G_\lambda x - G_\lambda y\|^2 - (\lambda - 2\rho)\langle G_\lambda x - G_\lambda y, Fx - Fy\rangle\\ &\quad + \langle Fx - Fy, x - y\rangle - \rho\|Fx - Fy\|^2\\ &= (\lambda - \rho)\|G_\lambda x - G_\lambda y\|^2 - (\lambda - 2\rho)\langle G_\lambda x - G_\lambda y, Fx - Fy\rangle\\ &\quad + \tfrac{L}{\hat{L}}\langle Fx - Fy, x - y\rangle + \tfrac{\hat{L} - L}{\hat{L}}\langle Fx - Fy, x - y\rangle - \rho\|Fx - Fy\|^2.\end{aligned}$$

Since F is $\frac{1}{L}$-co-coercive by our assumption, the last inequality leads to

$$\begin{aligned}\langle G_\lambda x - G_\lambda y, x - y\rangle &\geq (\lambda - \rho)\|G_\lambda x - G_\lambda y\|^2 + \big(\tfrac{1}{\hat{L}} - \rho\big)\|Fx - Fy\|^2\\ &\quad - (\lambda - 2\rho)\langle G_\lambda x - G_\lambda y, Fx - Fy\rangle + \tfrac{\hat{L} - L}{\hat{L}}\langle Fx - Fy, x - y\rangle\\ &= \tfrac{4(1 - \hat{L}\rho)(\lambda - \rho) - (\lambda - 2\rho)^2\hat{L}}{4(1 - \hat{L}\rho)}\|G_\lambda x - G_\lambda y\|^2 + \tfrac{\hat{L} - L}{\hat{L}}\langle Fx - Fy, x - y\rangle\\ &\quad + \tfrac{1 - \hat{L}\rho}{\hat{L}}\big\|Fx - Fy - \tfrac{(\lambda - 2\rho)\hat{L}}{2(1 - \hat{L}\rho)}(G_\lambda x - G_\lambda y)\big\|^2\\ &\geq \tfrac{\lambda(4 - \hat{L}\lambda) - 4\rho}{4(1 - \hat{L}\rho)}\|G_\lambda x - G_\lambda y\|^2 + \tfrac{\hat{L} - L}{\hat{L}}\langle Fx - Fy, x - y\rangle,\end{aligned}$$

which exactly proves (10), where $\bar{\beta} := \frac{\lambda(4 - \hat{L}\lambda) - 4\rho}{4(1 - \hat{L}\rho)} \geq 0$ and $\Lambda := \frac{\hat{L} - L}{L\hat{L}}$, provided that $\rho\hat{L} < 1$ and $\rho < \lambda \leq \frac{2 + 2\sqrt{1 - \hat{L}\rho}}{\hat{L}}$.

(ii) To prove (11), we denote by $x := J_{\lambda T}u$ and by $y := J_{\lambda T}v$ for given $u, v \in \mathrm{dom}(T)$, where $\lambda > \rho$. Then, we have $A_\lambda u := \frac{1}{\lambda}(u - J_{\lambda T}u) = \frac{1}{\lambda}(u - x)$. Now, by (8), we have $A_\lambda u = S_\lambda u - Fx$. Similarly, we also have $A_\lambda v = S_\lambda v - Fy$. Using these two relations and the $(\lambda - \rho)$-co-coercivity of A_λ, we can show that

$$\langle S_\lambda u - S_\lambda v - (Fx - Fy), u - v\rangle \geq (\lambda - \rho)\|S_\lambda u - S_\lambda v - (Fx - Fy)\|^2.$$

Utilizing again $x - \lambda Fx = u - \lambda S_\lambda u$ from (8), the last inequality leads to

$$\begin{aligned}\langle S_\lambda u - S_\lambda v, u - v\rangle &\geq (\lambda - \rho)\|S_\lambda u - S_\lambda v\|^2 + (\lambda - \rho)\|Fx - Fy\|^2\\ &\quad - (\lambda - 2\rho)\langle S_\lambda u - S_\lambda v, Fx - Fy\rangle + \langle Fx - Fy, u - v - \lambda(S_\lambda u - S_\lambda v)\rangle\\ &= (\lambda - \rho)\|S_\lambda u - S_\lambda v\|^2 - (\lambda - 2\rho)\langle S_\lambda u - S_\lambda v, Fx - Fy\rangle\\ &\quad + \tfrac{L}{\hat{L}}\langle Fx - Fy, x - y\rangle + \tfrac{\hat{L} - L}{\hat{L}}\langle Fx - Fy, x - y\rangle - \rho\|Fx - Fy\|^2.\end{aligned}$$

Substituting $\langle Fx - Fy, x - y\rangle \geq \frac{1}{L}\|Fx - Fy\|^2$ from the $\frac{1}{L}$-co-coercivity of F into the last inequality, we can further derive that

$$\begin{aligned}\langle S_\lambda u - S_\lambda v, u - v\rangle &\geq (\lambda-\rho)\|S_\lambda u - S_\lambda v\|^2 + \big(\tfrac{1}{\hat{L}} - \rho\big)\|Fx - Fy\|^2\\ &\quad - (\lambda - 2\rho)\langle S_\lambda u - S_\lambda v, Fx - Fy\rangle + \tfrac{\hat{L}-L}{\hat{L}}\langle Fx - Fy, x - y\rangle\\ &= \tfrac{4(1-\hat{L}\rho)(\lambda-\rho)-(\lambda-2\rho)^2\hat{L}}{4(1-\hat{L}\rho)}\|S_\lambda u - S_\lambda v\|^2 + \tfrac{(\hat{L}-L)}{\hat{L}}\langle Fx - Fy, x - y\rangle\\ &\quad + \tfrac{1-\hat{L}\rho}{\hat{L}}\|Fx - Fy - \tfrac{(\lambda-2\rho)\hat{L}}{2(1-\hat{L}\rho)}(S_\lambda u - S_\lambda v)\|^2\\ &\geq \tfrac{\lambda(4-\hat{L}\lambda)-4\rho}{4(1-\hat{L}\rho)}\|S_\lambda u - S_\lambda v\|^2 + \tfrac{\hat{L}-L}{\hat{L}}\langle Fx - Fy, x - y\rangle.\end{aligned}$$

This proves (11) with $\bar{\beta} := \frac{\lambda(4-\hat{L}\lambda)-4\rho}{4(1-\hat{L}\rho)} \geq 0$ and $\Lambda := \frac{\hat{L}-L}{L\hat{L}} \geq 0$ as in Statement (ii).

(iii) Finally, since $J_{\lambda T}x = x - \lambda A_{\lambda T}x$ and $J_{\lambda T}y = y - \lambda A_{\lambda T}y$, using the $(\lambda-\rho)$-co-coercivity of $A_{\lambda T}$, we can show that

$$\begin{aligned}\|J_{\lambda T}x - J_{\lambda T}y\|^2 &= \|x - y - \lambda(A_{\lambda T}x - A_{\lambda T}y)\|^2\\ &= \|x-y\|^2 - 2\lambda\langle A_{\lambda T}x - A_{\lambda T}y, x - y\rangle + \lambda^2\|A_{\lambda T}x - A_{\lambda T}y\|^2\\ &\leq \|x-y\|^2 - \lambda\big(\lambda - 2\rho\big)\|A_{\lambda T}x - A_{\lambda T}y\|^2.\end{aligned}$$

Thus, if $\lambda \geq 2\rho$, then $\|J_{\lambda T}x - J_{\lambda T}y\| \leq \|x - y\|$, implying that $J_{\lambda T}$ is nonexpansive. ■

Appendix B. The Proof of Technical Results in Section 3

We provide the full proof of all technical lemmas in the main text of Section 3.

B.1 The proof of Lemma 5: The L-SVRG estimator

Proof At the k-th iteration, we have 3 independent random variables: a mini-batch $\mathcal{S}_k$, a mega-batch $\bar{\mathcal{S}}_k$ to form $\bar{F}\tilde{x}^k$, and a Bernoulli's random variable i_k following the rule (13).

Define $\widehat{F}^k := F\tilde{x}^k + \mathbf{F}(x^k, \mathcal{S}_k) - \mathbf{F}(\tilde{x}^k, \mathcal{S}_k)$. Then, from (L-SVRG), we have $\widetilde{F}^k = \widehat{F}^k + \bar{F}\tilde{x}^k - F\tilde{x}^k$. By Young's inequality, for any $\tau > 0$, we can show that

$$\mathbb{E}_{(\mathcal{S}_k,\bar{\mathcal{S}}_k)}\big[\|\widetilde{F}^k - Fx^k\|^2\big] \leq (1+\tau)\mathbb{E}_{\mathcal{S}_k}\big[\|\widehat{F}^k - Fx^k\|^2\big] + \tfrac{1+\tau}{\tau}\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\|\bar{F}\tilde{x}^k - F\tilde{x}^k\|^2\big]. \tag{52}$$

For the expectation setting (E), we consider $X^k(\xi) := \mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^k,\xi) - (Fx^k - F\tilde{x}^k)$. Then, we have $\mathbb{E}_\xi\big[X^k(\xi)\big] = 0$. Since $\mathcal{S}_k$ is i.i.d., we can show that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_k}\big[\|\widehat{F}^k - Fx^k\|^2\big] &= \mathbb{E}_{\mathcal{S}_k}\big[\|F\tilde{x}^k + \mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(\tilde{x}^k,\mathcal{S}_k) - Fx^k\|^2\big]\\ &= \mathbb{E}_{\mathcal{S}_k}\big[\|\tfrac{1}{b_k}\textstyle\sum_{\xi_i\in\mathcal{S}_k}\big[\mathbf{F}(x^k,\xi_i) - \mathbf{F}(\tilde{x}^k,\xi_i) - (Fx^k - F\tilde{x}^k)\big]\|^2\big]\\ &= \mathbb{E}_{\mathcal{S}_k}\big[\|\tfrac{1}{b_k}\textstyle\sum_{\xi_i\in\mathcal{S}_k} X^k(\xi_i)\|^2\big]\\ &= \tfrac{1}{b_k}\mathbb{E}_\xi\big[\|X^k(\xi)\|^2\big]\\ &\leq \tfrac{1}{b_k}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^k,\xi)\|^2\big].\end{aligned}$$

For the finite-sum case (F), we denote by $X_i^k := F_ix^k - F_i\tilde{x}^k - (Fx^k - F\tilde{x}^k)$. Then, $\mathbb{E}_i\big[X_i^k\big] = 0$. Similar to Pham et al. (2020, Lemma 2), we can show that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}_k}\big[\|\widehat{F}^k - Fx^k\|^2\big] &= \mathbb{E}_{\mathcal{S}_k}\big[\|F\tilde{x}^k + \mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(\tilde{x}^k,\mathcal{S}_k) - Fx^k\|^2\big]\\ &= \mathbb{E}_{\mathcal{S}_k}\big[\|\tfrac{1}{b_k}\textstyle\sum_{i\in\mathcal{S}_k}\big[F_ix^k - F_i\tilde{x}^k - (Fx^k - F\tilde{x}^k)\big]\|^2\big]\\ &= \mathbb{E}_{\mathcal{S}_k}\big[\|\tfrac{1}{b_k}\textstyle\sum_{i\in\mathcal{S}_k}X_i^k\|^2\big]\\ &\le \tfrac{n-b_k}{(n-1)b_k}\cdot\tfrac{1}{n}\textstyle\sum_{i=1}^n\|F_ix^k - F_i\tilde{x}^k\|^2\\ &\le \tfrac{1}{b_kn}\textstyle\sum_{i=1}^n\|F_ix^k - F_i\tilde{x}^k\|^2\\ &= \tfrac{1}{b_k}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^k,\xi)\|^2\big], \quad \text{where } \mathbf{F}(x,\xi) := F_ix.\end{aligned}$$

Combining either the first or second relation above and (52), and then taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides of the result, we get

$$\mathbb{E}_k\big[\|\widetilde{F}^k - Fx^k\|^2\big] \le \tfrac{1+\tau}{b_k}\mathbb{E}_k\big[\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^k,\xi)\|^2\big]\big] + \tfrac{1+\tau}{\tau}\mathbb{E}_k\big[\|\bar{F}\tilde{x}^k - F\tilde{x}^k\|^2\big]. \tag{53}$$

If we define $\hat{\Delta}_k := \frac{1}{b_k}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^k,\xi)\|^2\big]$, then (53) implies (14).

Next, for a Bernoulli's random variable i_k following the rule (13), we have

$$\begin{aligned}\mathbb{E}_{\xi,i_k}\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^k,\xi)\|^2\big] &= \mathbf{p}_k\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big]\\ &\quad + (1-\mathbf{p}_k)\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^{k-1},\xi)\|^2\big].\end{aligned}$$

Now, for any $c > 0$, by Young's inequality, we have

$$\begin{aligned}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(\tilde{x}^{k-1},\xi)\|^2\big] &\le (1+c)\mathbb{E}_\xi\big[\|\mathbf{F}(x^{k-1},\xi) - \mathbf{F}(\tilde{x}^{k-1},\xi)\|^2\big]\\ &\quad + \big(1+\tfrac{1}{c}\big)\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big].\end{aligned}$$

Combining the last two expressions, taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides of the result, and using the definition of $\hat{\Delta}_k$ and $b_{k-1} \le b_k$, we can show that

$$\mathbb{E}_k\big[\hat{\Delta}_k\big] \le (1+c)(1-\mathbf{p}_k)\hat{\Delta}_{k-1} + \tfrac{1}{b_k}\big[(1-\mathbf{p}_k)\big(1+\tfrac{1}{c}\big) + \mathbf{p}_k\big]\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big].$$

Let us choose $c := \frac{(1-\alpha)\mathbf{p}_k}{1-\mathbf{p}_k}$ for some $\alpha \in (0,1)$. Then, we get $(1+c)(1-\mathbf{p}_k) = 1 - \alpha\mathbf{p}_k$ and $(1-\mathbf{p}_k)\big(1+\frac{1}{c}\big) + \mathbf{p}_k = \frac{1-(1+\alpha)\mathbf{p}_k+\mathbf{p}_k^2}{(1-\alpha)\mathbf{p}_k} \le \frac{1}{(1-\alpha)\mathbf{p}_k}$ for any $0 \le \mathbf{p}_k \le 1$. Hence, we obtain

$$\mathbb{E}_k\big[\hat{\Delta}_k\big] \le (1-\alpha\mathbf{p}_k)\hat{\Delta}_{k-1} + \tfrac{1}{(1-\alpha)b_k\mathbf{p}_k}\cdot\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big],$$

which proves (15).

Next, since $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\|\bar{F}\tilde{x}^k - F\tilde{x}^k\|^2\big] \le \frac{\sigma^2}{n_k}$, (53) implies

$$\mathbb{E}_k\big[\|\widetilde{F}^k - Fx^k\|^2\big] \le (1+\tau)\mathbb{E}_k\big[\hat{\Delta}_k\big] + \tfrac{(1+\tau)\sigma^2}{\tau n_k}. \tag{54}$$

Let us define $\Delta_k := (1+\tau)\hat{\Delta}_k + \frac{(1+\tau)\sigma^2}{\tau n_k}$. Then, plugging Δ_k from (54) into (15) and using $n_{k-1} \le n_k$, we can show that

$$\mathbb{E}_k\big[\Delta_k\big] \le (1-\alpha\mathbf{p}_k)\Delta_{k-1} + \tfrac{1+\tau}{(1-\alpha)b_k\mathbf{p}_k}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big] + \tfrac{(1+\tau)\alpha\mathbf{p}_k\sigma^2}{\tau n_k}. \tag{55}$$

If we choose $\tau := 1$, then using (54) and (55), we can show that $\widetilde{F}^k$ satisfies Definition 4 with $\Delta_k := 2\hat{\Delta}_k + \frac{2\sigma^2}{n_k}$, $\kappa_k := \alpha \mathbf{p}_k$, $\Theta_k := \frac{2}{(1-\alpha)b_k \mathbf{p}_k}$, and $\sigma_k^2 := \frac{2\alpha \mathbf{p}_k \sigma^2}{n_k}$.

Finally, if $\bar{F}\tilde{x}^k = F\tilde{x}^k$, then by setting $\tau = 0$ in (14) and combining the result and (15), they imply that $\widetilde{F}^k$ satisfies Definition 4 with $\Delta_k = \hat{\Delta}_k$, $\kappa_k = \alpha \mathbf{p}_k$, $\Theta_k = \frac{1}{(1-\alpha)b_k \mathbf{p}_k}$, and $\sigma_k^2 = 0$. ■

B.2 The proof of Lemma 6: The SAGA estimator

Proof Let $X_i^k := F_i x^k - \hat{F}_i^k$ for all $i \in [n]$. Then, we have $\mathbb{E}_i\big[X_i^k\big] = Fx^k - \frac{1}{n}\sum_{j=1}^n \hat{F}_j^k$ for any $i \in [n]$. Therefore, we can derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_k}\big[\|\widetilde{F}^k - Fx^k\|^2\big] &= \mathbb{E}_{\mathcal{S}_k}\big[\|\tfrac{1}{b_k}\textstyle\sum_{i\in\mathcal{S}_k} X_i^k - \big[Fx^k - \frac{1}{n}\sum_{j=1}^n \hat{F}_j^k\big]\|^2\big] \\
&= \mathbb{E}_{\mathcal{S}_k}\big[\|\tfrac{1}{b_k}\textstyle\sum_{i\in\mathcal{S}_k}\big(X_i^k - \mathbb{E}_i\big[X_i^k\big]\big)\|^2\big] \\
&= \tfrac{1}{b_k^2}\textstyle\sum_{i\in\mathcal{S}_k}\mathbb{E}_i\big[\|X_i^k - \mathbb{E}_i\big[X_i^k\big]\|^2\big] \\
&\leq \tfrac{1}{b_k^2}\textstyle\sum_{i\in\mathcal{S}_k}\mathbb{E}_i\big[\|X_i^k\|^2\big] \\
&= \tfrac{1}{nb_k}\textstyle\sum_{i=1}^n \|F_i x^k - \hat{F}_i^k\|^2.
\end{aligned}$$

Taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides of this inequality and using $\Delta_k := \frac{1}{nb_k}\sum_{i=1}^n \|F_i x^k - \hat{F}_i^k\|^2$, we obtain the first line of (17).

Now, from the definition of Δ_k and the update rule (16), for any $c > 0$, by Young's inequality, we can show that

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_k}\big[\Delta_k\big] &= \tfrac{1}{nb_k}\textstyle\sum_{i=1}^n \mathbb{E}_{\mathcal{S}_k}\big[\|F_i x^k - \hat{F}_i^k\|^2\big] \\
&\overset{(16)}{=} \big(1 - \tfrac{b_k}{n}\big)\tfrac{1}{nb_k}\textstyle\sum_{i=1}^n \|F_i x^k - \hat{F}_i^{k-1}\|^2 + \tfrac{b_k}{n}\cdot\tfrac{1}{nb_k}\textstyle\sum_{i=1}^n \|F_i x^k - F_i x^{k-1}\|^2 \\
&\leq \tfrac{(1+c)b_{k-1}}{b_k}\big(1 - \tfrac{b_k}{n}\big)\tfrac{1}{nb_{k-1}}\textstyle\sum_{i=1}^n \|F_i x^{k-1} - \hat{F}_i^{k-1}\|^2 \\
&\quad + \tfrac{(1+c)}{cnb_k}\big(1 - \tfrac{b_k}{n}\big)\textstyle\sum_{i=1}^n \|F_i x^k - F_i x^{k-1}\|^2 + \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \|F_i x^k - F_i x^{k-1}\|^2 \\
&= \tfrac{(1+c)b_{k-1}}{b_k}\big(1 - \tfrac{b_k}{n}\big)\Delta_{k-1} + \big[\tfrac{1}{n} + \big(1 - \tfrac{b_k}{n}\big)\tfrac{(1+c)}{cb_k}\big]\tfrac{1}{n}\textstyle\sum_{i=1}^n \|F_i x^k - F_i x^{k-1}\|^2.
\end{aligned}$$

For any $\alpha \in (0,1)$, if we choose $c := \frac{(n-\alpha b_k)b_k}{(n-b_k)b_{k-1}} - 1$, then $\frac{(1+c)b_{k-1}}{b_k}(1 - \frac{b_k}{n}) = 1 - \frac{\alpha b_k}{n}$. In addition, we can also compute $\underline{C}_k := \frac{1}{n} + \big(1 - \frac{b_k}{n}\big)\frac{(1+c)}{cb_k}$ as $\underline{C}_k = \frac{1}{n} + \frac{(n-b_k)(n-\alpha b_k)}{n[n(b_k - b_{k-1}) + b_k b_{k-1} - \alpha b_k^2]}$. Hence, taking $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides, we obtain from the last inequality that

$$\mathbb{E}_k\big[\Delta_k\big] \leq \big(1 - \tfrac{\alpha b_k}{n}\big)\Delta_{k-1} + \tfrac{\underline{C}_k}{n}\textstyle\sum_{i=1}^n \|F_i x^k - F_i x^{k-1}\|^2. \tag{56}$$

Since $b_{k-1} \geq b_k \geq b_{k-1} - \frac{(1-\alpha)b_k b_{k-1}}{2n}$, we have

$$n(b_k - b_{k-1}) + b_k b_{k-1} - \alpha b_k^2 \geq b_k b_{k-1} - \alpha b_k^2 - \tfrac{(1-\alpha)}{2} b_k b_{k-1} = \tfrac{1+\alpha}{2} b_k b_{k-1} - \alpha b_k^2 \geq \tfrac{(1-\alpha)b_k^2}{4}.$$

This implies that $\underline{C}_k \leq \frac{1}{n} + \frac{2(n-b_k)(n-\alpha b_k)}{n(1-\alpha)b_k^2} \leq \frac{1}{n} + \frac{2n}{(1-\alpha)b_k^2} \leq \frac{(3-\alpha)n}{(1-\alpha)b_k^2} =: \Theta_k$. Using this bound, we obtain from (56) the second bound in (17).

Consequently, $\widetilde{F}^k$ satisfies the $\mathbf{VR}(\kappa_k, \Theta_k, \Delta_k, \sigma_k)$ property in Definition 4 with $\kappa_k := \frac{\alpha b_k}{n} \in (0,1]$, Δ_k and Θ_k given above, and $\sigma_k^2 = 0$. ■

B.3 The proof of Lemma 7: The L-SARAH estimator

Proof We prove for the expectation setting (E). The proof of the finite-sum setting (F) is similar to the ones in Driggs et al. (2020); Li et al. (2020). Let i_k be the Bernoulli's random variable following the switching rule in (L-SARAH). Then, we have

$$\mathbb{E}_{i_k}\big[\|\widetilde{F}^k - Fx^k\|^2\big] = (1-\mathbf{p}_k)\|\widetilde{F}^{k-1} + \mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(x^{k-1},\mathcal{S}_k) - Fx^k\|^2 + \mathbf{p}_k\|\bar{F}x^k - Fx^k\|^2.$$

By the proof of the loopless SARAH estimator (Li et al., 2020, Lemma 3), we have

$$\begin{aligned} A_k &:= \mathbb{E}_k\big[\|\widetilde{F}^{k-1} + \mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(x^{k-1},\mathcal{S}_k) - Fx^k\|^2\big] \\ &= \|\widetilde{F}^{k-1} - Fx^{k-1}\|^2 + \mathbb{E}_k\big[\|\mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(x^{k-1},\mathcal{S}_k)\|^2\big] - \|Fx^k - Fx^{k-1}\|^2 \\ &\le \|\widetilde{F}^{k-1} - Fx^{k-1}\|^2 + \tfrac{1}{b_k}\mathbb{E}_\xi\big[\|\mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi)\|^2\big]. \end{aligned}$$

Taking the conditional expectation $\mathbb{E}_k\big[\cdot\big]$ on both sides of the first estimate and combining the result with the second expression, we obtain (18).

Finally, note that $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\|\bar{F}x^k - Fx^k\|^2\big] \le \frac{\sigma^2}{n_k}$ and $1-\mathbf{p}_k \le 1$, (18) shows that $\widetilde{F}^k$ satisfies Definition 4 with $\Delta_k := \|\widetilde{F}^k - Fx^k\|^2$, $\kappa_k = \mathbf{p}_k$, $\Theta_k := \frac{1}{b_k}$, and $\sigma_k^2 := \frac{\mathbf{p}_k\sigma^2}{n_k}$. However, if we choose $\bar{F}x^k = Fx^k$, then we can set $\sigma_k = 0$ since $\mathbb{E}_{\bar{\mathcal{S}}_k}\big[\|\bar{F}x^k - Fx^k\|^2\big] = 0$. ■

B.4 The proof of Lemma 8: The HSGD estimator

Proof Let $e^k := \widetilde{F}^k - Fx^k$, $X^k(\xi) := \mathbf{F}(x^k,\xi) - \mathbf{F}(x^{k-1},\xi) - (Fx^k - Fx^{k-1})$, and $Y^k := \bar{F}x^k - Fx^k$. Then, we have $\mathbb{E}_\xi\big[X^k(\xi)\big] = 0$ and $\mathbb{E}_{\hat{\mathcal{S}}_k}\big[Y^k\big] = 0$ by our assumption. In addition, if we denote $X^k := \frac{1}{b_k}\sum_{\xi\in\mathcal{S}_k} X^k(\xi)$, then we also have $\mathbb{E}_{\mathcal{S}_k}\big[\|X^k\|^2\big] \le \frac{1}{b_k}\bar{\mathcal{E}}_k$ for $\bar{\mathcal{E}}_k$ defined in Definition 4. From (HSGD), we can write

$$\begin{aligned} e^k &= \widetilde{F}^k - Fx^k = (1-\tau_k)\widetilde{F}^{k-1} + (1-\tau_k)[\mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(x^{k-1},\mathcal{S}_k)] + \tau_k\bar{F}x^k - Fx^k \\ &= (1-\tau_k)(\widetilde{F}^{k-1} - Fx^{k-1}) + (1-\tau_k)[\mathbf{F}(x^k,\mathcal{S}_k) - \mathbf{F}(x^{k-1},\mathcal{S}_k) - (Fx^k - Fx^{k-1})] \\ &\quad + \tau_k(\bar{F}x^k - Fx^k) \\ &= (1-\tau_k)e^{k-1} + (1-\tau_k)X^k + \tau_k Y^k. \end{aligned}$$

This expression leads to

$$\begin{aligned} \|e^k\|^2 &= (1-\tau_k)^2\|e^{k-1}\|^2 + (1-\tau_k)^2\|X^k\|^2 + \tau_k^2\|Y^k\|^2 \\ &\quad + 2(1-\tau_k)^2\langle e^{k-1}, X^k\rangle + 2(1-\tau_k)\tau_k\langle X^k, Y^k\rangle + 2(1-\tau_k)\tau_k\langle e^{k-1}, Y^k\rangle. \end{aligned}$$

Taking $\mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[\cdot\big]$ on both sides of this expression and using $\mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[X^k\big] = \mathbb{E}_{\hat{\mathcal{S}}_k}\big[\mathbb{E}_{\mathcal{S}_k}\big[X^k \mid \hat{\mathcal{S}}_k\big]\big] = 0$ and $\mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[Y^k\big] = \mathbb{E}_{\mathcal{S}_k}\big[\mathbb{E}_{\hat{\mathcal{S}}_k}\big[Y^k \mid \mathcal{S}_k\big]\big] = 0$, we can show that

$$\begin{aligned} \mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[\|e^k\|^2\big] &= (1-\tau_k)^2\|e^{k-1}\|^2 + (1-\tau_k)^2\mathbb{E}_{\mathcal{S}_k}\big[\|X^k\|^2\big] + \tau_k^2\mathbb{E}_{\hat{\mathcal{S}}_k}\big[\|Y^k\|^2\big] \\ &\quad + 2(1-\tau_k)\tau_k\mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[\langle X^k, Y^k\rangle\big]. \end{aligned} \tag{57}$$

Here, we have used the facts that X^k only depends on $\mathcal{S}_k$ and Y^k only depends on $\hat{\mathcal{S}}_k$. Now, we consider two cases as follows.

(i) If $\mathcal{S}_k$ and $\hat{\mathcal{S}}_k$ are independent, then $\mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[\langle X^k, Y^k\rangle \mid \mathcal{F}_k\big] = 0$. Using this fact, $\mathbb{E}_{\mathcal{S}_k}\big[\|X^k\|^2\big] \le \frac{1}{b_k}\bar{\mathcal{E}}_k$, and $\delta_k^2 := \mathbb{E}_{\hat{\mathcal{S}}_k}\big[\|Y^k\|^2\big]$ into (57), and then taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides of the result, we obtain (19).

(ii) If $\mathcal{S}_k$ and $\hat{\mathcal{S}}_k$ are not independent, then by Young's inequality, we have

$$2(1-\tau_k)\tau_k\mathbb{E}_{(\mathcal{S}_k,\hat{\mathcal{S}}_k)}\big[\langle X^k, Y^k\rangle\big] \le (1-\tau_k)^2\mathbb{E}_{\mathcal{S}_k}\big[\|X^k\|^2\big] + \tau_k^2\mathbb{E}_{\hat{\mathcal{S}}_k}\big[\|Y^k\|^2\big].$$

Substituting this inequality, $\mathbb{E}_{\mathcal{S}_k}\big[\|X^k\|^2\big] \le \frac{1}{b_k}\bar{\mathcal{E}}_k$, and $\delta_k^2 := \mathbb{E}_{\hat{\mathcal{S}}_k}\big[\|Y^k\|^2\big]$ into (57), and taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides of the result, we obtain (20).

Finally, to prove (iii), we utilize $\mathbb{E}_{\hat{\mathcal{S}}_k}\big[\delta_k^2\big] \le \frac{\sigma^2}{\hat{b}_k}$ to obtain (12) in Definition 4 from (19) and (20), respectively. ■

Appendix C. The Proof of Technical Results in Section 4

This appendix presents the full proof of technical lemmas and theorems in Section 4.

C.1 The proof of Lemma 9: Lower bounding the difference $\mathcal{L}_k - \mathcal{L}_{k+1}$

Proof From the first two lines of (VFOSA$_+$), we can easily show that $t_kx^{k+1} = (t_k-1)x^k + z^k - t_k\eta_k\widetilde{G}_\lambda^k$. Rearranging this expression and using $z^k = z^{k+1} - \nu(x^{k+1}-y^k) = z^{k+1} + \nu\eta_k\widetilde{G}_\lambda^k$ from the last line of (VFOSA$_+$), we obtain

$$\left\{\begin{array}{rcl} t_k(t_k-1)(x^{k+1}-x^k) &=& -(t_k-1)(x^k-z^k) - t_k(t_k-1)\eta_k\widetilde{G}_\lambda^k,\\ t_k(t_k-1)(x^{k+1}-x^k) &=& -t_k(x^{k+1}-z^k) - t_k^2\eta_k\widetilde{G}_\lambda^k\\ &=& -t_k(x^{k+1}-z^{k+1}) - t_k(t_k-\nu)\eta_k\widetilde{G}_\lambda^k. \end{array}\right. \tag{58}$$

Let us define $\mathcal{E}_{k+1} := L\langle Fx^{k+1} - Fx^k, x^{k+1}-x^k\rangle$ as in Lemma 9. Then, from (10), we have

$$\begin{aligned}\widetilde{\mathcal{T}}_{[1]} &:= t_k(t_k-1)\langle G_\lambda x^{k+1}, x^{k+1}-x^k\rangle - t_k(t_k-1)\langle G_\lambda x^k, x^{k+1}-x^k\rangle\\ &\ge \beta t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 + (\bar{\beta}-\beta)t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\\ &\quad + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}.\end{aligned}$$

Substituting (58) into $\widetilde{\mathcal{T}}_{[1]}$ and using $e_\lambda^k := \widetilde{G}_\lambda^k - G_\lambda x^k$ and $\theta_k := \frac{t_k-1}{t_k-\nu}$, we can show that

$$\begin{aligned}\widehat{\mathcal{T}}_{[1]} &:= (t_k-1)\langle G_\lambda x^k, x^k - z^k\rangle - t_k\langle G_\lambda x^{k+1}, x^{k+1}-z^{k+1}\rangle\\ &\ge \eta_k t_k(t_k-\nu)\langle G_\lambda x^{k+1}, \widetilde{G}_\lambda^k\rangle - \eta_k t_k(t_k-1)\langle G_\lambda x^k, \widetilde{G}_\lambda^k\rangle + \beta t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\\ &\quad + (\bar{\beta}-\beta)t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}\\ &= \eta_k t_k(t_k-\nu)\langle G_\lambda x^{k+1}, G_\lambda x^k\rangle - \eta_k t_k(t_k-1)\|G_\lambda x^k\|^2 + \beta t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\\ &\quad + (\bar{\beta}-\beta)t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}\\ &\quad + \eta_k t_k(t_k-\nu)\langle G_\lambda x^{k+1} - \theta_k G_\lambda x^k, e_\lambda^k\rangle.\end{aligned}$$

By Young's inequality, for any $s > 0$, we can derive from $\widehat{\mathcal{T}}_{[1]}$ that

$$\begin{aligned}\widehat{\mathcal{T}}_{[1]} &:= (t_k-1)\langle G_\lambda x^k, x^k - z^k\rangle - t_k\langle G_\lambda x^{k+1}, x^{k+1} - z^{k+1}\rangle\\ &\ge \eta_k t_k(t_k-\nu)\langle G_\lambda x^{k+1}, G_\lambda x^k\rangle - \eta_k t_k(t_k-1)\|G_\lambda x^k\|^2 + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}\\ &\quad + \beta t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 + (\bar{\beta}-\beta)t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\\ &\quad - s\beta t_k(t_k-\nu)\|G_\lambda x^{k+1} - \theta_k G_\lambda x^k\|^2 - \tfrac{\eta_k^2 t_k(t_k-\nu)}{4s\beta}\|e_\lambda^k\|^2.\end{aligned}$$

Substituting the following identity

$$\begin{aligned}\|G_\lambda x^{k+1} - \theta_k G_\lambda x^k\|^2 &= (1-\theta_k)\|G_\lambda x^{k+1}\|^2 - \theta_k(1-\theta_k)\|G_\lambda x^k\|^2 + \theta_k\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\\ &= \tfrac{1-\nu}{t_k-\nu}\|G_\lambda x^{k+1}\|^2 - \tfrac{(1-\nu)(t_k-1)}{(t_k-\nu)^2}\|G_\lambda x^k\|^2 + \tfrac{t_k-1}{t_k-\nu}\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\end{aligned}$$

into the last expression $\widehat{\mathcal{T}}_{[1]}$, one can show that

$$\begin{aligned}\mathcal{T}_{[1]} &:= t_{k-1}\langle G_\lambda x^k, x^k - z^k\rangle - t_k\langle G_\lambda x^{k+1}, x^{k+1} - z^{k+1}\rangle\\ &\ge \beta t_k\big[t_k - 1 - s(1-\nu)\big]\|G_\lambda x^{k+1}\|^2 - t_k(t_k-1)\big[\eta_k - \beta - \tfrac{s\beta(1-\nu)}{t_k-\nu}\big]\|G_\lambda x^k\|^2\\ &\quad + t_k\big[\eta_k(t_k-\nu) - 2\beta(t_k-1)\big]\langle G_\lambda x^{k+1}, G_\lambda x^k\rangle + (t_{k-1} - t_k + 1)\langle G_\lambda x^k, x^k - z^k\rangle\\ &\quad + \big[\bar{\beta} - (1+s)\beta\big]t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 + \Lambda t_k(t_k-1)\mathcal{E}_{k+1} - \tfrac{\eta_k^2 t_k(t_k-\nu)}{4s\beta}\|e_\lambda^k\|^2.\end{aligned}$$

Since $z^{k+1} - z^k = -\nu\eta_k\widetilde{G}_\lambda^k$, by Young's inequality again, we have

$$\begin{aligned}\mathcal{T}_{[2]} &:= \tfrac{1-\mu}{2\nu\eta_k}\|z^k - x^\star\|^2 - \tfrac{1-\mu}{2\nu\eta_{k+1}}\|z^{k+1} - x^\star\|^2\\ &= \tfrac{1-\mu}{2\nu\eta_k}\big[\|z^k - x^\star\|^2 - \|z^{k+1} - x^\star\|^2\big] + \tfrac{(1-\mu)}{2\nu}\big(\tfrac{1}{\eta_k} - \tfrac{1}{\eta_{k+1}}\big)\|z^{k+1} - x^\star\|^2\\ &= -\tfrac{1-\mu}{\nu\eta_k}\langle z^{k+1} - z^k, z^k - x^\star\rangle - \tfrac{1-\mu}{2\nu\eta_k}\|z^{k+1} - z^k\|^2 + \tfrac{(1-\mu)}{2\nu}\big(\tfrac{1}{\eta_k} - \tfrac{1}{\eta_{k+1}}\big)\|z^{k+1} - x^\star\|^2\\ &= (1-\mu)\langle\widetilde{G}_\lambda^k, z^k - x^\star\rangle - \tfrac{(1-\mu)\nu\eta_k}{2}\|\widetilde{G}_\lambda^k\|^2 + \tfrac{(1-\mu)}{2\nu}\big(\tfrac{1}{\eta_k} - \tfrac{1}{\eta_{k+1}}\big)\|z^{k+1} - x^\star\|^2\\ &= (1-\mu)\langle G_\lambda x^k, z^k - x^\star\rangle + (1-\mu)\langle e_\lambda^k, z^k - x^\star\rangle - \tfrac{(1-\mu)\nu\eta_k}{2}\|\widetilde{G}_\lambda^k\|^2\\ &\quad + \tfrac{(1-\mu)}{2\nu}\big(\tfrac{1}{\eta_k} - \tfrac{1}{\eta_{k+1}}\big)\|z^{k+1} - x^\star\|^2\\ &\ge (1-\mu)\langle G_\lambda x^k, z^k - x^\star\rangle - \tfrac{(1-\mu)\nu\beta(t_{k-1}-1)(t_k-1)}{\mu(1-\nu)}\|e_\lambda^k\|^2 - \tfrac{\mu(1-\mu)(1-\nu)}{4\nu\beta(t_{k-1}-1)(t_k-1)}\|z^k - x^\star\|^2\\ &\quad - (1-\mu)\nu\eta_k\|G_\lambda x^k\|^2 - (1-\mu)\nu\eta_k\|e_\lambda^k\|^2 + \tfrac{(1-\mu)}{2\nu}\big(\tfrac{1}{\eta_k} - \tfrac{1}{\eta_{k+1}}\big)\|z^{k+1} - x^\star\|^2.\end{aligned}$$

Since $\eta_k = \frac{2\beta(t_k-1)}{t_k-\nu}$ and $t_k = \mu(k+r)$ due to (27), we have $\eta_k(t_k-\nu) - 2\beta(t_k-1) = 0$ and $t_{k-1} - t_k + 1 = 1 - \mu$, respectively. In addition, we also have $\frac{1}{\eta_k} - \frac{1}{\eta_{k+1}} = \frac{\mu(1-\nu)}{2\beta(t_k-1)(t_{k+1}-1)}$.

Adding $\mathcal{T}_{[2]}$ to $\mathcal{T}_{[1]}$ and then using the last three relations of parameters, we can derive that

$$\begin{aligned}
\mathcal{T}_{[3]} &:= t_{k-1}\langle G_\lambda x^k, x^k - z^k\rangle - t_k\langle G_\lambda x^{k+1}, x^{k+1} - z^{k+1}\rangle\\
&\quad + \tfrac{1-\mu}{2\nu\eta_k}\|z^k - x^\star\|^2 - \tfrac{1-\mu}{2\nu\eta_{k+1}}\|z^{k+1} - x^\star\|^2\\
&\geq \beta t_k\big[t_k - 1 - s(1-\nu)\big]\|G_\lambda x^{k+1}\|^2\\
&\quad - \tfrac{\beta(t_k-1)}{t_k-\nu}\big[t_k(t_k - 2 + \nu - s(1-\nu)) + 2(1-\mu)\nu\big]\|G_\lambda x^k\|^2\\
&\quad + \big[\bar{\beta} - (1+s)\beta\big]t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}\\
&\quad - \big[\tfrac{\beta t_k(t_k-1)^2}{s(t_k-\nu)} + \tfrac{(1-\mu)\nu\beta(t_{k-1}-1)(t_k-1)}{\mu(1-\nu)} + \tfrac{2\beta\nu(1-\mu)(t_k-1)}{t_k-\nu}\big]\|e_\lambda^k\|^2\\
&\quad - \tfrac{\mu(1-\mu)(1-\nu)}{4\nu\beta(t_{k-1}-1)(t_k-1)}\|z^k - x^\star\|^2 + \tfrac{\mu(1-\mu)(1-\nu)}{4\nu\beta(t_k-1)(t_{k+1}-1)}\|z^{k+1} - x^\star\|^2\\
&\quad + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle.
\end{aligned}$$

Since $\frac{1}{\eta_k} + \frac{\mu(1-\nu)}{2\beta(t_{k-1}-1)(t_k-1)} = \frac{(t_k-\nu)(t_{k-1}-1)+\mu(1-\nu)}{2\beta(t_{k-1}-1)(t_k-1)}$, using $\mathcal{L}_k$ from (25) and $\|e_\lambda^k\| \leq \|\widetilde{F}^k - Fx^k\| = \|e^k\|$ from (22), the last inequality leads to

$$\begin{aligned}
\mathcal{L}_k - \mathcal{L}_{k+1} \geq\ & \beta\varphi_k\cdot\|G_\lambda x^k\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle + \Lambda t_k(t_k-1)\mathcal{E}_{k+1}\\
&+ \big[\bar{\beta} - (1+s)\beta\big]t_k(t_k-1)\|G_\lambda x^{k+1} - G_\lambda x^k\|^2 - \psi_k\|e^k\|^2,
\end{aligned}$$

which proves (28), where φ_k and ψ_k are respectively

$$\begin{aligned}
\varphi_k &:= t_{k-1}\big[t_{k-1} - 1 - s(1-\nu)\big] - \tfrac{(t_k-1)}{t_k-\nu}\big[t_k(t_k - 2 + \nu - s(1-\nu)) + 2\nu(1-\mu)\big],\\
\psi_k &:= \beta(t_k-1)\big[\tfrac{t_k(t_k-1)}{s(t_k-\nu)} + \tfrac{2\nu(1-\mu)}{t_k-\nu} + \tfrac{\nu(1-\mu)(t_{k-1}-1)}{\mu(1-\nu)}\big],
\end{aligned}$$

as given in (29). ■

C.2 The proof of Lemma 10: The lower bound of $\mathcal{L}_k$

Proof First, since F is $\frac{1}{L}$-co-coercive, it is monotone. From (10), $G_\lambda x^\star = 0$ for any $x^\star \in \operatorname{zer}(\Phi)$, and the monotonicity of F, we have $\langle G_\lambda x^k, x^k - x^\star\rangle \geq \bar{\beta}\|G_\lambda x^k\|^2$.

Next, utilizing the last inequality, (25), and $\eta_k = \frac{2\beta(t_k-1)}{t_k-\nu}$ from (27), we can show that

$$\begin{aligned}
\mathcal{L}_k &:= \beta a_k\|G_\lambda x^k\|^2 + t_{k-1}\langle G_\lambda x^k, x^k - z^k\rangle + \tfrac{(1-\mu)[(t_k-\nu)(t_{k-1}-1)+\mu(1-\nu)]}{4\nu\beta(t_{k-1}-1)(t_k-1)}\|z^k - x^\star\|^2\\
&= \tfrac{\beta}{2}\big(2a_k - t_{k-1}^2\big)\|G_\lambda x^k\|^2 + \big[\tfrac{(1-\mu)[(t_k-\nu)(t_{k-1}-1)+\mu(1-\nu)]}{4\nu\beta(t_{k-1}-1)(t_k-1)} - \tfrac{1}{2\beta}\big]\|z^k - x^\star\|^2\\
&\quad + \tfrac{\beta t_{k-1}^2}{2}\|G_\lambda x^k\|^2 + t_{k-1}\langle G_\lambda x^k, x^k - x^\star\rangle - t_{k-1}\langle G_\lambda x^k, z^k - x^\star\rangle + \tfrac{1}{2\beta}\|z^k - x^\star\|^2\\
&= \tfrac{\beta(2a_k - t_{k-1}^2)}{2}\|G_\lambda x^k\|^2 + \tfrac{(t_{k-1}-1)[(1-\mu-2\nu)t_k + \nu(1+\mu)] + \mu(1-\mu)(1-\nu)}{4\nu\beta(t_{k-1}-1)(t_k-1)}\|z^k - x^\star\|^2\\
&\quad + t_{k-1}\langle G_\lambda x^k, x^k - x^\star\rangle + \tfrac{1}{2\beta}\|z^k - x^\star - \beta t_{k-1}G_\lambda x^k\|^2\\
&\geq \tfrac{\beta(2a_k - t_{k-1}^2) + 2\bar{\beta}t_{k-1}}{2}\|G_\lambda x^k\|^2 + \tfrac{(t_{k-1}-1)[(1-\mu-2\nu)t_k + \nu(1+\mu)] + \mu(1-\mu)(1-\nu)}{4\nu\beta(t_{k-1}-1)(t_k-1)}\|z^k - x^\star\|^2.
\end{aligned}$$

Now, we have $A_k := \beta(2a_k - t_{k-1}^2) + 2\bar{\beta}t_{k-1} = \beta t_{k-1}[t_{k-1} - 2 - 2s(1-\nu)] + 2\bar{\beta}t_{k-1}$. Using this relation into the last estimate, we obtain (30).

Finally, from the definitions of $\mathcal{P}_k$ and $\mathcal{Q}_k$ in (25), the nonnegativity of the last terms in $\mathcal{P}_k$ and $\mathcal{Q}_k$, and Assumption 1.2, we can easily show that $\mathcal{P}_k \geq \mathcal{Q}_k \geq \mathcal{L}_k \geq 0$. ■

C.3 The proof of Lemma 11: Lower bounding the term $\mathcal{P}_k - \mathbb{E}_k\big[\mathcal{P}_{k+1}\big]$

Proof First, taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ on both sides of (28), we obtain

$$\begin{aligned}\mathcal{L}_k \;\geq\; & \mathbb{E}_k\big[\mathcal{L}_{k+1}\big] + \beta\varphi_k\|G_\lambda x^k\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle \\ & + \big[\bar{\beta} - (1+s)\beta\big]t_k(t_k-1)\mathbb{E}_k\big[\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\big] \\ & + \Lambda t_k(t_k-1)\mathbb{E}_k\big[\mathcal{E}_{k+1}\big] - \psi_k\mathbb{E}_k\big[\|e^k\|^2\big].\end{aligned}$$

Here, ψ_k and φ_k from (29) are respectively given by

$$\begin{aligned}\psi_k &:= \beta(t_k-1)\big[\tfrac{t_k(t_k-1)}{s(t_k-\nu)} + \tfrac{2\nu(1-\mu)}{t_k-\nu} + \tfrac{\nu(1-\mu)(t_{k-1}-1)}{\mu(1-\nu)}\big], \\ \varphi_k &:= t_{k-1}\big[t_{k-1} - 1 - s(1-\nu)\big] - \tfrac{(t_k-1)}{t_k-\nu}\big[t_k(t_k - 2 + \nu - s(1-\nu)) + 2(1-\mu)\nu\big].\end{aligned}$$

Adding $\big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 + \Lambda t_{k-1}(t_{k-1}-1)\mathcal{E}_k$ to both sides of the last inequality, and using $\mathcal{Q}_k$ from (25) we have

$$\begin{aligned}\mathcal{Q}_k \;\geq\; & \mathbb{E}_k\big[\mathcal{Q}_{k+1}\big] + \beta\varphi_k\|G_\lambda x^k\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle \\ & + \big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 \\ & + \Lambda t_{k-1}(t_{k-1}-1)\mathcal{E}_k - \psi_k\mathbb{E}_k\big[\|e^k\|^2\big].\end{aligned} \tag{59}$$

Next, from (22) and Definition 4, we have

$$\mathbb{E}_k\big[\|e^k\|^2\big] \;=\; \mathbb{E}_k\big[\|\widetilde{F}^k - Fx^k\|^2\big] \leq \mathbb{E}_k\big[\Delta_k\big]. \tag{60}$$

Let $\bar{\mathcal{E}}_k$ be defined in Definition 4. Then, by Assumption 1.2, we have

$$\mathcal{E}_k \;=\; L\cdot\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle \geq \bar{\mathcal{E}}_k. \tag{61}$$

Now, from (12) in Definition 4 and (60), for $\Gamma_k \geq 0$ in (25), we can show that

$$\begin{aligned}\tfrac{(1-\kappa_k)\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\Delta_{k-1} \;\geq\; & \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\mathbb{E}_k\big[\Delta_k\big] - \tfrac{\Theta_k\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k - \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2 \\ \geq\; & \psi_k\mathbb{E}_k\big[\|e^k\|^2\big] + \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1) - 2\psi_k}{2}\mathbb{E}_k\big[\Delta_k\big] - \tfrac{\Theta_k\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k \\ & - \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2.\end{aligned}$$

Adding this inequality to (59), and then using (61) we get

$$\begin{aligned}\mathcal{T}_{[1]} \;:=\; & \mathcal{Q}_k + \tfrac{(1-\kappa_k)\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\Delta_{k-1} \\ \geq\; & \mathbb{E}_k\big[\mathcal{Q}_{k+1}\big] + \tfrac{[\Gamma_k t_{k-1}(t_{k-1}-1) - 2\psi_k]}{2}\mathbb{E}_k\big[\Delta_k\big] + \beta\varphi_k\|G_\lambda x^k\|^2 \\ & + \big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle \\ & + \tfrac{t_{k-1}(t_{k-1}-1)}{2}\big(2\Lambda - \Theta_k\Gamma_k\big)\bar{\mathcal{E}}_k - \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2.\end{aligned} \tag{62}$$

Assume that $[\Gamma_{k+1}(1-\kappa_{k+1})+\mu^{-1}\beta]t_k(t_k-1) \le \Gamma_k t_{k-1}(t_{k-1}-1) - 2\psi_k$, which is exactly the condition (31). Then, using $\mathcal{P}_k$ from (25) and this condition, we obtain from (62) that

$$\begin{aligned}
\mathcal{P}_k &:= \mathcal{Q}_k + \tfrac{[\mu(1-\kappa_k)\Gamma_k+\beta]t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1}\\
&\ge \mathbb{E}_k\big[\mathcal{Q}_{k+1}\big] + \tfrac{[\mu(1-\kappa_{k+1})\Gamma_{k+1}+\beta]t_k(t_k-1)}{2\mu}\mathbb{E}_k\big[\Delta_k\big] + \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1}\\
&\quad + \beta\varphi_k\|G_\lambda x^k\|^2 + (1-\mu)\langle G_\lambda x^k, x^k - x^\star\rangle\\
&\quad + \big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2\\
&\quad + \tfrac{t_{k-1}(t_{k-1}-1)}{2}\big(2\Lambda - \Gamma_k\Theta_k\big)\bar{\mathcal{E}}_k - \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2.
\end{aligned}$$

Finally, using again $\mathcal{P}_{k+1} := \mathcal{Q}_{k+1} + \frac{[\mu(1-\kappa_{k+1})\Gamma_{k+1}+\beta]t_k(t_k-1)}{2\mu}\Delta_k$ from (25), the last expression implies (32). ■

C.4 The proof of Theorem 12: Key bounds

Proof Suppose that we choose $\nu := \frac{\mu}{2}$, $s := \frac{2(\mu-\nu)}{1-\nu} = \frac{2\mu}{2-\mu}$, and $0 < \beta \le \frac{\bar{\beta}}{1+s} = \frac{(2-\mu)\bar{\beta}}{2+\mu}$. Using these choices and $t_k := \mu(k+r)$ from (33) for $r \ge 2 + \frac{1}{\mu}$, φ_k and ψ_k defined by (29) respectively become

$$\begin{aligned}
\varphi_k &:= t_{k-1}(t_{k-1} - 1 - 2(\mu-\nu)) - \tfrac{t_k-1}{t_k-\nu}\big[t_k^2 - (2+2\mu-3\nu)t_k + 2(1-\mu)\nu\big]\\
&\ge \tfrac{(2-3\mu)t_k}{2} + 3\mu^2,\\
\psi_k &:= \beta(t_k-1)\big[\tfrac{(1-\nu)t_k(t_k-1)}{2(\mu-\nu)(t_k-\nu)} + \tfrac{2(1-\mu)\nu}{t_k-\nu} + \tfrac{(1-\mu)\nu(t_{k-1}-1)}{\mu(1-\nu)}\big]\\
&\le \tfrac{2\beta}{\mu}t_k(t_k-1).
\end{aligned}$$

If we choose $0 < \mu < \frac{2}{3}$ as in (33), then from the first expression, we get $\varphi_k > 0$.

From the second expression, we can check that (31) of Lemma 11 holds if we impose

$$\big[\Gamma_{k+1}(1-\kappa_{k+1}) + \tfrac{5\beta}{\mu}\big]t_k(t_k-1) \le \Gamma_k t_{k-1}(t_{k-1}-1).$$

This condition is equivalent to $\kappa_{k+1} \ge 1 - \frac{\Gamma_k t_{k-1}(t_{k-1}-1)}{\Gamma_{k+1}t_k(t_k-1)} + \frac{5\beta}{\mu\Gamma_{k+1}}$, which is exactly the first condition in (34).

Using φ_k, we can derive from (32) that

$$\begin{aligned}
\mathcal{P}_k \ge\ & \mathbb{E}_k\big[\mathcal{P}_{k+1}\big] + \tfrac{\beta}{2}\big[(2-3\mu)t_k + 6\mu^2\big]\|G_\lambda x^k\|^2\\
&- \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2 + \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1} + \tfrac{(2\Lambda-\Gamma_k\Theta_k)t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k \qquad (63)\\
&+ \big[\bar{\beta} - \tfrac{(2+\mu)\beta}{2-\mu}\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2.
\end{aligned}$$

Let us denote by $\hat{\psi}_k := \beta\big[(2-3\mu)t_k + 6\mu^2\big] \ge 0$. Then, (63) is equivalent to

$$\begin{aligned}
\mathbb{E}_k\big[\mathcal{P}_{k+1}\big] \le\ & \mathcal{P}_k - \tfrac{\hat{\psi}_k}{2}\|G_\lambda x^k\|^2 - \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1} + \tfrac{\Gamma_k t_{k-1}(t_{k-1}-1)}{2}\sigma_k^2\\
&- \big[\bar{\beta} - \tfrac{(2+\mu)\beta}{2-\mu}\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 \qquad (64)\\
&- \tfrac{(2\Lambda-\Gamma_k\Theta_k)t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k.
\end{aligned}$$

Since $2\Lambda - \Gamma_k\Theta_k \geq 0$ from the second condition of (34), dropping the last term of (64), and then taking the total expectation $\mathbb{E}[\cdot]$ on both sides of the result, we get

$$\begin{aligned}\mathbb{E}[\mathcal{P}_{k+1}] \leq \mathbb{E}[\mathcal{P}_k] - \tfrac{\hat{\psi}_k}{2}\mathbb{E}[\|G_\lambda x^k\|^2] - \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\mathbb{E}[\Delta_{k-1}] + \tfrac{t_{k-1}(t_{k-1}-1)}{2}\Gamma_k\sigma_k^2 \\ - \big[\bar{\beta} - \tfrac{(2+\mu)\beta}{2-\mu}\big]t_{k-1}(t_{k-1}-1)\mathbb{E}[\|G_\lambda x^k - G_\lambda x^{k-1}\|^2].\end{aligned} \tag{65}$$

Let $\underline{B}_K := \frac{1}{2\beta}\sum_{k=0}^K \Gamma_k t_{k-1}(t_{k-1}-1)\sigma_k^2$. Then, since $\Gamma_k \leq \frac{2\Lambda}{\Theta_k}$ by (34), we can show that

$$\underline{B}_K := \tfrac{1}{2\beta}\textstyle\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\Gamma_k\sigma_k^2 \leq B_K := \tfrac{\Lambda}{\beta}\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\tfrac{\sigma_k^2}{\Theta_k}.$$

By induction and the last relation, we obtain from (65) that

$$\begin{aligned}&\mathbb{E}[\mathcal{P}_K] && \leq \mathcal{P}_0 + \beta B_{K-1}\\ &\textstyle\sum_{k=0}^K \hat{\psi}_k\mathbb{E}[\|G_\lambda x^k\|^2] && \leq 2\mathbb{E}[\mathcal{P}_0] + 2\beta B_K,\\ &\textstyle\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\mathbb{E}[\Delta_{k-1}] && \leq \tfrac{2\mu}{\beta}\big(\mathbb{E}[\mathcal{P}_0] + \beta B_K\big),\\ &\textstyle\sum_{k=0}^K \big[\bar{\beta} - \tfrac{(2+\mu)\beta}{2-\mu}\big]t_{k-1}(t_{k-1}-1)\mathbb{E}[\|G_\lambda x^k - G_\lambda x^{k-1}\|^2] && \leq \mathbb{E}[\mathcal{P}_0] + \beta B_K.\end{aligned} \tag{66}$$

Because $x^0 = z^0$, we have

$$\begin{aligned}\mathcal{L}_0 &:= \beta a_0\|G_\lambda x^0\|^2 + \tfrac{1-\mu}{2\nu\eta_0}\|x^0 - x^\star\|^2\\ &\leq \beta\mu(r-1)(\mu r - 2\mu - 1)\|G_\lambda x^0\|^2 + \tfrac{2r-1}{4\beta(\mu r-1)}\|x^0 - x^\star\|^2\\ &\leq \beta\mathcal{R}_0^2, \quad \text{where } \mathcal{R}_0^2 := \mu^2 r^2\|G_\lambda x^0\|^2 + \tfrac{2r-1}{4\beta^2(\mu r-1)}\|x^0 - x^\star\|^2.\end{aligned}$$

Since $\kappa_0 \in (0,1]$, $x^{-1} = x^0$, $\Delta_{-1} = \Delta_0$, and $\mathbb{E}[\Delta_0] \leq \sigma_0^2$, from (25), we get

$$\begin{aligned}\mathbb{E}[\mathcal{P}_0] &= \mathbb{E}[\mathcal{Q}_0] = \mathbb{E}[\mathcal{L}_0] + \tfrac{(r-1)(\mu r-\mu-1)[(1-\kappa_0)\mu\Gamma_0+\beta]}{2}\mathbb{E}[\Delta_0]\\ &\leq \beta\big[\mathbb{E}[\mathcal{R}_0^2] + \tfrac{\mu r^2(\mu\Gamma_0+\beta)}{2\beta}\sigma_0^2\big]\\ &= \beta\big(\Psi_0^2 + E_0^2\big),\end{aligned}$$

where $\Psi_0^2 := \mathbb{E}[\mathcal{R}_0^2]$ and $E_0^2 := \frac{\mu r^2(\mu\Gamma_0+\beta)}{2\beta}\sigma_0^2$ are given in (36).

Now, since $A_k := \beta t_{k-1}(t_{k-1} - 2 - 2\mu) + 2\bar{\beta}t_{k-1} \geq \beta[t_{k-1}^2 - 2(1+\mu)t_{k-1} + \frac{2(2+\mu)}{2-\mu}t_{k-1}] \geq \beta t_{k-1}^2$ and $1 - \mu - 2\nu \geq 0$, from Lemma 10, we also have

$$\mathcal{L}_k \geq \tfrac{A_k}{2}\|G_\lambda x^k\|^2 \geq \tfrac{\beta t_{k-1}^2}{2}\|G_\lambda x^k\|^2 = \tfrac{\beta\mu^2(k+r-1)^2}{2}\|G_\lambda x^k\|^2.$$

Combining the last two bounds and the first estimate of (66), we can derive that

$$\tfrac{\beta\mu^2(K+r-1)^2}{2}\mathbb{E}[\|G_\lambda x^K\|^2] \leq \mathbb{E}[\mathcal{L}_K] \leq \mathbb{E}[\mathcal{P}_K] \leq \beta\big(\Psi_0^2 + E_0^2 + B_K\big).$$

This expression leads to (35).

Next, since $\hat{\psi}_k = \beta\big[(2-3\mu)t_k + 6\mu^2\big] > 0$ due to $0 < \mu < \frac{2}{3}$, we obtain from the second estimate of (66) that

$$\textstyle\sum_{k=0}^K \beta\big[(2-3\mu)\mu(k+r) + 6\mu^2\big]\mathbb{E}[\|G_\lambda x^k\|^2] \leq 2\mathbb{E}[\mathcal{P}_0] + 2\beta B_K.$$

Substituting $\mathbb{E}\big[\mathcal{P}_0\big] \le \beta(\Psi_0^2 + E_0^2)$ into this inequality, we obtain the first line of (37). The second bound of (37) is a consequence of the third line of (66) and the facts that $\mathbb{E}\big[\|e^{k-1}\|^2\big] \le \mathbb{E}\left[\Delta_{k-1}\right]$ and $\mathbb{E}\big[\mathcal{P}_0\big] \le \beta(\Psi_0^2 + E_0^2)$. Similarly, the third line of (37) is a direct consequence of the last line of (66) and $\mathbb{E}\big[\mathcal{P}_0\big] \le \beta(\Psi_0^2 + E_0^2)$. ■

C.5 The proof of Theorem 13: The $\mathcal{O}\left(1/k^2\right)$ and $o\big(1/k^2\big)$-convergence rates

Proof Let us divide this proof into several steps as follows.

Step 1. Proving (39)***.*** Since $B_\infty := \frac{\Lambda}{\beta}\sum_{k=0}^{\infty} t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} < +\infty$ by (38), (39) follows directly from (35) and $B_k \le B_\infty$.

Step 2. The first two summability bounds of (40)***.*** By our condition (38), we have $\Psi_0^2 + B_K \le \Psi_0^2 + B_\infty$. Combining this fact, $e_\lambda^k := \widetilde{G}_\lambda^k - G_\lambda x^k$, $e^k := \widetilde{F}^k - Fx^k$ and $\|e_\lambda^k\| \le \|e^k\|$ from (22), we can show from (37) that

$$\begin{aligned}
&\textstyle\sum_{k=0}^{\infty} t_k\mathbb{E}\big[\|G_\lambda x^k\|^2\big] &&< +\infty,\\
&\textstyle\sum_{k=0}^{\infty} t_k^2\mathbb{E}\big[\|e_\lambda^k\|^2\big] &&< +\infty,\\
&\textstyle\sum_{k=0}^{\infty} t_k^2\mathbb{E}\big[\|e^k\|^2\big] &&< +\infty,\\
&\textstyle\sum_{k=0}^{\infty} t_k^2\mathbb{E}\big[\|G_\lambda x^{k+1} - G_\lambda x^k\|^2\big] &&< +\infty.
\end{aligned} \tag{67}$$

These prove the first and second lines of (40). Here, the last line of (67) requires $0 < \beta < \frac{(2-\mu)\bar{\beta}}{2+\mu}$. Combining $\|\widetilde{G}_\lambda^k\|^2 \le 2\|e_\lambda^k\|^2 + 2\|G_\lambda x^k\|^2$ and the first two lines of (67) we get.

$$\textstyle\sum_{k=0}^{\infty} t_k\mathbb{E}\big[\|\widetilde{G}_\lambda^k\|^2\big] < +\infty. \tag{68}$$

Step 3. The last summability bound of (40)***.*** From (VFOSA$_+$) we can show that

$$\begin{aligned}
t_k(x^{k+1} - x^k + \eta_k\widetilde{G}_\lambda^k) &= z^k - x^k,\\
(t_{k-1}-1)(x^k - x^{k-1} + \eta_{k-1}\widetilde{G}_\lambda^{k-1}) &= z^{k-1} - x^k - \eta_{k-1}\widetilde{G}_\lambda^{k-1}\\
&= z^k - x^k - (1-\nu)\eta_{k-1}\widetilde{G}_\lambda^{k-1}.
\end{aligned}$$

Combining these two expressions, we get

$$t_k(x^{k+1} - x^k + \eta_k\widetilde{G}_\lambda^k) = (t_{k-1}-1)(x^k - x^{k-1} + \eta_{k-1}\widetilde{G}_\lambda^{k-1}) + (1-\nu)\eta_{k-1}\widetilde{G}_\lambda^{k-1}. \tag{69}$$

If we denote $v^k := x^{k+1} - x^k + \eta_k\widetilde{G}_\lambda^k$, then we can rewrite the last expression as

$$t_k v^k := \big(1 - \tfrac{1}{t_{k-1}}\big)t_{k-1}v^{k-1} + \tfrac{(1-\nu)\eta_{k-1}t_{k-1}}{t_{k-1}}\widetilde{G}_\lambda^{k-1}.$$

By convexity of $\|\cdot\|^2$ and $\frac{1}{t_{k-1}} \in (0,1]$, since $\eta_k = \frac{2\beta(t_k-1)}{t_k-\nu} \le 2\beta$, we have

$$\begin{aligned}
t_k^2\|v^k\|^2 &\le \big(1 - \tfrac{1}{t_{k-1}}\big)t_{k-1}^2\|v^{k-1}\|^2 + (1-\nu)^2\eta_{k-1}^2 t_{k-1}\|\widetilde{G}_\lambda^{k-1}\|^2\\
&\le t_{k-1}^2\|v^{k-1}\|^2 - t_{k-1}\|v^{k-1}\|^2 + 4(1-\nu)^2\beta^2 t_{k-1}\|\widetilde{G}_\lambda^{k-1}\|^2.
\end{aligned} \tag{70}$$

Taking the total expectation of this inequality, and then applying Lemma 28 with the fact that $\sum_{k=0}^{\infty} t_k\mathbb{E}\big[\|\widetilde{G}_\lambda^k\|^2\big] < +\infty$ from (68), we conclude that

$$\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|v^k\|^2\big] \text{ exists} \quad \text{and} \quad \sum_{k=0}^{\infty} t_k\mathbb{E}\big[\|v^k\|^2\big] < +\infty. \tag{71}$$

By Young's inequality and $\eta_k \le 2\beta$, we have

$$\|x^{k+1} - x^k\|^2 \le 2\|x^{k+1} - x^k + \eta_k\widetilde{G}_\lambda^k\|^2 + 2\eta_k^2\|\widetilde{G}_\lambda^k\|^2 \le 2\|v^k\|^2 + 8\beta^2\|\widetilde{G}_\lambda^k\|^2.$$

Combining this inequality, (68), (71), we get the last summability bound of (40).

Step 4. The limit of $t_k^2\mathbb{E}\big[\|v^k\|^2\big]$**.** Since $\sum_{k=0}^{\infty} t_k\mathbb{E}\big[\|v^k\|^2\big] < +\infty$ and $\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|v^k\|^2\big]$ exists, applying Lemma 29, we conclude that

$$\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|v^k\|^2\big] = \lim_{k\to\infty} k^2\mathbb{E}\big[\|v^k\|^2\big] = 0. \tag{72}$$

Step 5. The first limit of (41)**.** From (69), we can show that

$$\begin{aligned} t_k[x^{k+1} - x^k + \eta_k(\widetilde{G}_\lambda^k - G_\lambda x^k)] &= (t_{k-1} - 1)\big[x^k - x^{k-1} + \eta_{k-1}(\widetilde{G}_\lambda^{k-1} - G_\lambda x^{k-1})\big] \\ &\quad + (1-\nu)\eta_{k-1}\widetilde{G}_\lambda^{k-1} - t_k\eta_k G_\lambda x^k + (t_{k-1} - 1)\eta_{k-1}G_\lambda x^{k-1}. \end{aligned}$$

Let us define $w^k := x^{k+1} - x^k + \eta_k e_\lambda^k$ for $e_\lambda^k := \widetilde{G}_\lambda^k - G_\lambda x^k$. Then this expression becomes

$$\begin{aligned} t_k w^k &= (t_{k-1} - 1)w^{k-1} + (1-\nu)\eta_{k-1}\widetilde{G}_\lambda^{k-1} - t_k\eta_k G_\lambda x^k + (t_{k-1} - 1)\eta_{k-1}G_\lambda x^{k-1} \\ &= \big(1 - \tfrac{1}{t_{k-1}}\big)t_{k-1}\big[w^{k-1} - \tfrac{t_k\eta_k}{t_{k-1}-1}(G_\lambda x^k - G_\lambda x^{k-1})\big] \\ &\quad + \tfrac{1}{t_{k-1}}\Big[t_{k-1}[\eta_{k-1}(t_{k-1} - 1) - t_k\eta_k]G_\lambda x^{k-1} + (1-\nu)t_{k-1}\eta_{k-1}e_\lambda^{k-1}\Big]. \end{aligned}$$

By convexity of $\|\cdot\|^2$, Young's inequality, $\bar{\beta}$-co-coercivity of G_λ from (10), and $\|e_\lambda^k\| \le \|e^k\|$ from (22), we can deduce that

$$\begin{aligned} t_k^2\|w^k\|^2 &\le (t_{k-1} - 1)\,t_{k-1}\|w^{k-1} - \tfrac{t_k\eta_k}{t_{k-1}-1}(G_\lambda x^k - G_\lambda x^{k-1})\|^2 \\ &\quad + t_{k-1}\|[\eta_{k-1}(t_{k-1} - 1) - t_k\eta_k]G_\lambda x^{k-1} + (1-\nu)\eta_{k-1}e_\lambda^{k-1}\|^2 \\ &\le (t_{k-1} - 1)\,t_{k-1}\|w^{k-1}\|^2 + \tfrac{t_{k-1}t_k^2\eta_k^2}{t_{k-1}-1}\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 \\ &\quad - 2t_{k-1}t_k\eta_k\langle G_\lambda x^k - G_\lambda x^{k-1}, x^k - x^{k-1}\rangle - 2t_{k-1}t_k\eta_{k-1}\eta_k\langle G_\lambda x^k - G_\lambda x^{k-1}, e_\lambda^{k-1}\rangle \\ &\quad + 2t_{k-1}[\eta_{k-1}(t_{k-1} - 1) - t_k\eta_k]^2\|G_\lambda x^{k-1}\|^2 + 2(1-\nu)^2t_{k-1}\eta_{k-1}^2\|e_\lambda^{k-1}\|^2 \\ &\le (t_{k-1} - 1)\,t_{k-1}\|w^{k-1}\|^2 + t_{k-1}t_k\eta_k\Big(\tfrac{t_k\eta_k}{t_{k-1}-1} - 2\bar{\beta} + \eta_{k-1}\Big)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 \\ &\quad + 2t_{k-1}(\eta_{k-1}(t_{k-1} - 1) - t_k\eta_k)^2\|G_\lambda x^{k-1}\|^2 \\ &\quad + t_{k-1}\eta_{k-1}[2(1-\nu)^2\eta_{k-1} + t_k\eta_k]\|e^{k-1}\|^2. \end{aligned}$$

Let us denote by $Z_k := s_k\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 + \hat{s}_k\|G_\lambda x^{k-1}\|^2 + \tilde{s}_k\|e^{k-1}\|^2$, where

$$\begin{aligned} s_k &:= t_{k-1}t_k\eta_k\big(\tfrac{t_k\eta_k}{t_{k-1}-1} - 2\bar{\beta} + \eta_{k-1}\big) &&\le 8\beta^2t_k^2, \\ \hat{s}_k &:= 2t_{k-1}[\eta_{k-1}(t_{k-1} - 1) - t_k\eta_k]^2 &&\le 8\beta^2(\mu+\nu)^2t_{k-1}, \\ \tilde{s}_k &:= t_{k-1}\eta_{k-1}[2(1-\nu)^2\eta_{k-1} + t_k\eta_k] &&\le 4\beta^2t_{k-1}(t_k + 2) \le 8\beta^2t_k^2. \end{aligned}$$

Then, the last expression can be briefly rewritten as

$$t_k^2\|w^k\|^2 \le t_{k-1}^2\|w^{k-1}\|^2 - t_{k-1}\|w^{k-1}\|^2 + Z_k. \tag{73}$$

Taking the total expectation $\mathbb{E}\big[\,\cdot\,\big]$ on both sides of this inequality, we get

$$t_k^2\mathbb{E}\big[\|w^k\|^2\big] \le t_{k-1}^2\mathbb{E}\big[\|w^{k-1}\|^2\big] - t_{k-1}\mathbb{E}\big[\|w^{k-1}\|^2\big] + \mathbb{E}\big[Z_k\big].$$

Using the upper bound of the coefficients s_k, $\hat{s}_k$, and $\tilde{s}_k$, and (67), we can easily show that $\sum_{k=0}^{\infty}\mathbb{E}\big[Z_k\big] < +\infty$. Applying again Lemma 28 to the last inequality, we conclude that

$$\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|w^k\|^2\big] \text{ exists} \quad \text{and} \quad \textstyle\sum_{k=0}^{\infty} t_k\mathbb{E}\big[\|w^k\|^2\big] < +\infty.$$

Applying Lemma 29, these statements imply that $\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|w^k\|^2\big] = 0$.

By Young's inequality and $\|e_\lambda^k\| \le \|e^k\|$, one has

$$\|x^{k+1} - x^k\|^2 \le 2\|x^{k+1} - x^k + \eta_k e_\lambda^k\|^2 + 2\eta_k^2\|e_\lambda^k\|^2 \le 2\|w^k\|^2 + 8\beta^2\|e^k\|^2. \tag{74}$$

Combining this fact, $\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|w^k\|^2\big] = 0$ and (67), we prove the first limit of (41).

Step 6. The second limit of (41)***.*** Finally, since $\eta_k = \frac{2\beta(t_k-1)}{t_k-\nu} \ge \frac{4\beta(r\mu-1)}{\mu(2r-1)}$, by Young's inequality, we have

$$\begin{aligned}\tfrac{16\beta^2(r\mu-1)^2}{\mu^2(2r-1)^2}\|G_\lambda x^k\|^2 &\le \eta_k^2\|G_\lambda x^k\|^2\\ &\le 2\|x^{k+1} - x^k + \eta_k(\widetilde{G}_\lambda^k - G_\lambda x^k)\|^2 + 2\|x^{k+1} - x^k + \eta_k\widetilde{G}_\lambda^k\|^2\\ &\le 2\|w^k\|^2 + 2\|v^k\|^2.\end{aligned}$$

This fact, (72), and $\lim_{k\to\infty} t_k^2\mathbb{E}\big[\|w^k\|^2\big] = 0$ imply the second limit of (41). ■

C.6 The proof of Theorem 14: Almost sure convergence

Proof Again, we divide this proof into the following steps.

Step 1. The first summability bound in (43)***.*** First, applying our assumption $\Gamma_k\Theta_k \le \Lambda$ from (42), we can deduce from (64) that

$$\begin{aligned}\mathbb{E}\big[\mathcal{P}_{k+1} \mid \mathcal{F}_k\big] \le\ & \mathcal{P}_k - \tfrac{\hat{\psi}_k}{2}\|G_\lambda x^k\|^2 - \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1} - \tfrac{\Lambda t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k\\ &- \big(\bar{\beta} - \tfrac{(2+\mu)\beta}{2-\mu}\big)t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 + \tfrac{\Lambda t_{k-1}(t_{k-1}-1)\sigma_k^2}{2\Theta_k}.\end{aligned} \tag{75}$$

Next, we denote $U_k := \mathcal{P}_k$, $\gamma_k := 0$, $E_k := \frac{\Lambda t_{k-1}(t_{k-1}-1)\sigma_k^2}{2\Theta_k}$, and

$$\begin{aligned}V_k :=\ & \tfrac{\hat{\psi}_k}{2}\|G_\lambda x^k\|^2 + \tfrac{\beta t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1} + (\bar{\beta} - 2\beta)t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2\\ &+ \tfrac{\Lambda t_{k-1}(t_{k-1}-1)}{2}\bar{\mathcal{E}}_k.\end{aligned}$$

These quantities are nonnegative. By (38), we have $\sum_{k=0}^{\infty} E_k = \frac{\beta}{2}B_\infty < +\infty$ surely. In addition, $\sum_{k=0}^{\infty}\gamma_k = 0 < +\infty$ and $\hat{\psi}_k = 2\beta[(2-5\mu)t_k - (2-3\mu-2\mu^2)] = \mathcal{O}\big(t_k\big) > 0$ surely. Moreover, $\{\mathcal{F}_k\}$ is a filtration. By Lemma 30, we conclude that

$$\begin{aligned}
&\lim_{k\to\infty}\mathcal{P}_k \quad \text{exists almost surely,}\\
&\textstyle\sum_{k=0}^{\infty} t_k\|G_\lambda x^k\|^2 &&< +\infty \quad \text{almost surely,}\\
&\textstyle\sum_{k=0}^{\infty} t_k^2\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 &&< +\infty \quad \text{almost surely,}\\
&\textstyle\sum_{k=0}^{\infty} t_k^2\Delta_k &&< +\infty \quad \text{almost surely,}\\
&\textstyle\sum_{k=0}^{\infty} t_k^2\bar{\mathcal{E}}_k &&< +\infty \quad \text{almost surely.}
\end{aligned} \tag{76}$$

The second line of (76) is the first summability bound of (43).

Step 2. The second summability bound in (43)**.** Since $\mathbb{E}_k\big[\|e^k\|^2\big] \le \mathbb{E}_k\big[\Delta_k\big]$ by the first line of (12) and $\Gamma_k\Theta_k \le \Lambda$ by (42), utilizing these relations and the second line of (12), we get

$$\begin{aligned}
\Gamma_k\mathbb{E}_k\big[\|e^k\|^2\big] &\le \Gamma_k\mathbb{E}_k\big[\Delta_k\big] \overset{(12)}{\le} \Gamma_k(1-\kappa_k)\Delta_{k-1} + \Gamma_k\Theta_k\bar{\mathcal{E}}_k + \Gamma_k\sigma_k^2\\
&\le \Gamma_k(1-\kappa_k)\Delta_{k-1} + \Lambda\bar{\mathcal{E}}_k + \tfrac{\Lambda\sigma_k^2}{\Theta_k}.
\end{aligned}$$

Multiplying both sides of this inequality by $t_{k-1}(t_{k-1}-1)$ and utilizing $(1-\kappa_k)\Gamma_k t_{k-1}(t_{k-1}-1) \le \Gamma_{k-1}t_{k-2}(t_{k-2}-1)$ from (31) and $\Gamma_{k-1} \le \frac{\Lambda}{\underline{\Theta}}$ from (42), we can show that

$$\begin{aligned}
\Gamma_k t_{k-1}(t_{k-1}-1)\mathbb{E}_k\big[\|e^k\|^2\big] &\le \Gamma_{k-1}t_{k-2}(t_{k-2}-1)\Delta_{k-1} + \Lambda t_{k-1}(t_{k-1}-1)\big(\bar{\mathcal{E}}_k + \tfrac{\sigma_k^2}{\Theta_k}\big)\\
&\le \Gamma_{k-1}t_{k-2}(t_{k-2}-1)\|e^{k-1}\|^2 - \Gamma_{k-1}t_{k-2}(t_{k-2}-1)\|e^{k-1}\|^2\\
&\quad + \tfrac{\Lambda}{\underline{\Theta}}t_{k-2}(t_{k-2}-1)\Delta_{k-1} + \Lambda t_{k-1}(t_{k-1}-1)\big(\bar{\mathcal{E}}_k + \tfrac{\sigma_k^2}{\Theta_k}\big).
\end{aligned}$$

Let us denote by

$$E_k := \tfrac{\Lambda}{\underline{\Theta}}t_{k-2}(t_{k-2}-1)\Delta_{k-1} + \Lambda t_{k-1}(t_{k-1}-1)\bar{\mathcal{E}}_k + \Lambda t_{k-1}(t_{k-1}-1)\tfrac{\sigma_k^2}{\Theta_k}.$$

Then, by (76) and (38), we have $\sum_{k=0}^{\infty}E_k \le \frac{\Lambda}{\underline{\Theta}}\sum_{k=0}^{\infty}t_k^2\Delta_k + \Lambda\sum_{k=0}^{\infty}t_k^2\bar{\mathcal{E}}_k + \beta B_\infty < +\infty$ almost surely. Thus, applying again Lemma 30 with $U_k := \Gamma_{k-1}t_{k-2}(t_{k-2}-1)\|e^{k-1}\|^2$, $\gamma_k := 0$, $V_k := \Gamma_{k-1}t_{k-2}(t_{k-2}-1)\|e^{k-1}\|^2$, and E_k given above, we conclude that

$$\textstyle\sum_{k=0}^{\infty}\Gamma_k t_k^2\|e^k\|^2 < +\infty \quad \text{almost surely.}$$

However, since $\Gamma_k \ge \underline{\Gamma} > 0$ by (42), this implies the second summability bound in (43).

Step 3. The last summability bound in (43)**.** Since $\|\widetilde{G}_\lambda^k\|^2 \le 2\|G_\lambda x^k\|^2 + 2\|e_\lambda^k\|^2 \le 2\|G_\lambda x^k\|^2 + 2\|e^k\|^2$, using the second lines of (76) and (43), this inequality implies that

$$\textstyle\sum_{k=0}^{\infty}t_k\|\widetilde{G}_\lambda^k\|^2 < +\infty \quad \text{almost surely.} \tag{77}$$

Let $v^k := x^{k+1} - x^k + \eta_k\widetilde{G}_\lambda^k$. Taking the conditional expectation $\mathbb{E}_k\big[\,\cdot\,\big]$ of (70), we get

$$\mathbb{E}_k\big[t_k^2\|v^k\|^2\big] \le t_{k-1}^2\|v^{k-1}\|^2 - t_{k-1}\|v^{k-1}\|^2 + 4(1-\nu)^2\beta^2 t_{k-1}\|\widetilde{G}_\lambda^{k-1}\|^2.$$

Since $\sum_{k=0}^{\infty}t_k\|\widetilde{G}_\lambda^k\|^2 < +\infty$ by (77), applying Lemma 30 to the last inequality, and then using Lemma 29, we conclude that the following statements hold almost surely:

$$\lim_{k\to\infty}t_k^2\|v^k\|^2 = 0 \quad \text{and} \quad \textstyle\sum_{k=0}^{\infty}t_k\|v^k\|^2 < +\infty. \tag{78}$$

Let $w^k := x^{k+1} - x^k + \eta_k e_\lambda^k$ as before. Following similar arguments here, but applying them to (73), we can also prove that the following statements hold almost surely:

$$\lim_{k\to\infty} t_k^2\|w^k\|^2 = 0 \quad \text{and} \quad \sum_{k=0}^{\infty} t_k\|w^k\|^2 < +\infty. \tag{79}$$

Combining (74), (79), and the second line of (43), we can easily prove that $\sum_{k=0}^{\infty} t_k\|x^{k+1} - x^k\|^2 < +\infty$ almost surely. This is the last line of (43).

Step 4. The limits of (44)**.** Combining the limits in (78) and (79) and Young's inequality, similar to **Step 6** in the proof of Theorem 13, we can easily show that

$$\lim_{k\to\infty} t_k^2\|G_\lambda x^k\|^2 = 0 \quad \text{almost surely.} \tag{80}$$

This limit is the first limit in (44). The second limit of (44) follows from (74), the first limit of (79), and $\lim_{k\to\infty} t_k^2\|e^k\|^2 = 0$ almost surely from (43).

Step 5. The existence of $\lim_{k\to\infty}\|z^k - x^\star\|^2$**.** From (VFOSA$_+$), we can easily shows that $z^k - x^k = t_k(x^{k+1} - x^k) + t_k\eta_k\widetilde{G}_\lambda^k = t_k v^k$. Since $\lim_{k\to\infty} t_k^2\|v^k\|^2 = 0$ almost surely as shown in (78), we conclude that $\lim_{k\to\infty}\|z^k - x^k\| = 0$ almost surely. Using this limit and $\lim_{k\to\infty} t_k^2\|G_\lambda x^k\|^2 = 0$ from (80), we have

$$2|t_k\langle G_\lambda x^k, x^k - z^k\rangle \le \|x^k - z^k\|^2 + t_k^2\|G_\lambda x^k\|^2 \to 0 \quad \text{as } k\to\infty \quad \text{almost surely.}$$

Hence, we get $\lim_{k\to\infty} t_{k-1}|\langle G_\lambda x^k, x^k - z^k\rangle| = 0$ almost surely.

Next, as discussed in Assumption 1.2, F is L-Lipschitz continuous, we have

$$|\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle| \le \|Fx^k - Fx^{k-1}\|\|x^k - x^{k-1}\| = L\|x^k - x^{k-1}\|^2.$$

Since $\lim_{k\to\infty} t_{k-1}^2\|x^k - x^{k-1}\|^2 = 0$ due to the second limit of (44), we conclude that $\lim_{k\to\infty} t_{k-1}(t_{k-1}-1)|\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle| = 0$ almost surely.

Now, from (25) we can write

$$\begin{aligned}\mathcal{P}_k \;=\; & \beta a_k\|G_\lambda x^k\|^2 + t_{k-1}\langle G_\lambda x^k, x^k - z^k\rangle + \tfrac{[\mu(1-\kappa_k)\Gamma_k+\beta]t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1}\\
& + \big[\bar{\beta} - \tfrac{(2+\mu)\beta}{2-\mu}\big]t_{k-1}(t_{k-1}-1)\|G_\lambda x^k - G_\lambda x^{k-1}\|^2\\
& + \Lambda L t_{k-1}(t_{k-1}-1)\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle + \tfrac{c_k}{2\nu\beta}\|z^k - x^\star\|^2.\end{aligned} \tag{81}$$

Collecting all the necessary almost sure limits proven above, we have, almost surely

$$\begin{aligned}
&\lim_{k\to\infty}\mathcal{P}_k \;\text{ exists as shown in (76)},\\
&\lim_{k\to\infty} a_k\|G_\lambda x^k\|^2 = \lim_{k\to\infty} t_k^2\|G_\lambda x^k\|^2 = 0 \quad \text{by (80)},\\
&\lim_{k\to\infty} t_{k-1}|\langle G_\lambda x^k, x^k - z^k\rangle| = 0,\\
&\lim_{k\to\infty} t_{k-1}(t_{k-1}-1)|\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle| = 0,\\
&\lim_{k\to\infty} t_k^2\|G_\lambda x^k - G_\lambda x^{k-1}\|^2 = 0 \quad \text{from (76)},\\
&\lim_{k\to\infty} t_k^2\Delta_k = 0 \quad \text{from (76)}.
\end{aligned}$$

Combining these limits and (81), and noting that $c_k = \mathcal{O}\big(1\big)$ due to (26), we conclude that $\lim_{k\to\infty}\|z^k - x^\star\|^2$ exists almost surely.

Step 6. Almost sure convergence of $\{x^k\}$ ***and*** $\{z^k\}$**.** Since $\lim_{k\to\infty}\|x^k - z^k\| = 0$ and $\lim_{k\to\infty}\|z^k - x^\star\|$ exists almost surely, combining these facts and $|\|x^k - x^\star\| - \|z^k - x^\star\|| \le \|x^k - z^k\|$, we conclude that $\lim_{k\to\infty}\|x^k - x^\star\|^2$ exists almost surely.

Finally, since $\lim_{k\to\infty}\|x^k - x^\star\|^2$ exists almost surely for any solution $x^\star \in \mathrm{zer}(G_\lambda)$, $\lim_{k\to\infty}\|G_\lambda x^k\| = 0$ almost surely by the second line of (76), and G_λ is continuous, applying Lemma 31, we conclude that $\{x^k\}$ converges to a random variable $\bar{x}^\star \in \mathrm{zer}(G_\lambda)$ almost surely. However, since $\bar{x}^\star \in \mathrm{zer}(G_\lambda)$ if and only if $\bar{x}^\star \in \mathrm{zer}(\Phi)$ surely, we also conclude that $\{x^k\}$ almost surely converges to a solution $\bar{x}^\star \in \mathrm{zer}(\Phi)$. In addition, since $\lim_{k\to\infty}\|z^k - x^k\| = 0$ almost surely, we can state that $\{z^k\}$ almost surely converges to $\bar{x}^\star \in \mathrm{zer}(\Phi)$. ■

Appendix D. The Proof of Corollaries in Subsections 4.4 and 4.5

We now present the full proofs of Corollaries in Subsections 4.4 and 4.5.

D.1 The finite-sum setting (F)

This appendix presents the proof of Corollaries 18, 19, and 20, respectively.

D.1.1 The proof of Corollary 18: The L-SVRG variant of VFOSA$_+$

Proof Since the L-SVRG estimator is used and $\bar{F}\tilde{x}^k := F\tilde{x}^k$, by Lemma 5, we have

$$\Delta_k = \hat{\Delta}_k, \quad \kappa_k = \alpha\mathbf{p}_k, \quad \Theta_k = \tfrac{1}{(1-\alpha)b_k\mathbf{p}_k} \ge \underline{\Theta} := \tfrac{1}{(1-\alpha)n} > 0, \quad \text{and} \quad \sigma_k^2 = 0.$$

Thus, the quantity B_K in Theorem 12 becomes $B_K := \frac{\Lambda}{\beta}\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} = 0$. Moreover, since $x^0 = \tilde{x}^0$, we also have $\Delta_0 = \hat{\Delta}_0 := \frac{1}{nb_0}\sum_{i=1}^n \|F_i x^0 - F_i\tilde{x}^0\|^2 \le \sigma_0^2 = 0$.

Next, let us choose $\alpha = \frac{1}{2}$ and assume that $\Gamma_k = \frac{5c_p\beta n^\omega}{\mu} =: \underline{\Gamma} > 0$ for all $k \ge 0$ and given $c_p > 0$ and $\omega > 0$. We recall the first condition of (34) in Theorem 12 as follows:

$$\begin{aligned}\kappa_k \;&=\; \alpha\mathbf{p}_k = \tfrac{\mathbf{p}_k}{2} \ge 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_{k-1}(t_{k-1}-1)} + \tfrac{5\beta}{\mu\Gamma_k} = 1 + \tfrac{5\beta}{\mu\Gamma_k} - \tfrac{t_{k-2}(t_{k-2}-1)}{t_{k-1}(t_{k-1}-1)}\\ &=\; \tfrac{1}{c_p n^\omega} + \tfrac{2\mu}{\mu(k+r-1)-1} - \tfrac{\mu+1}{(k+r-1)(\mu(k+r-1)-1)}.\end{aligned}$$

This condition holds if $\mathbf{p}_k \ge \frac{2}{c_p n^\omega} + \frac{4\mu}{\mu(k+r-1)-1}$.

Now, for $r > 5 + \frac{1}{\mu}$ and $n^\omega \ge \frac{1}{c_p}\max\left\{\frac{2\mu(r-1)-2}{\mu(r-5)-1}, \frac{\mu(r-1)-1}{4\mu}\right\}$, if we choose

$$\mathbf{p}_k := \begin{cases} \frac{2}{c_p n^\omega} + \frac{4\mu}{\mu(k+r-1)-1} & \text{if } 0 \le k \le K_0 := \lfloor 4c_p n^\omega + 1 + \mu^{-1} - r\rfloor,\\ \frac{3}{c_p n^\omega} & \text{otherwise,}\end{cases}$$

then $0 < \frac{2}{c_p n^\omega} + \frac{4\mu}{\mu(k+r-1)-1} \le \mathbf{p}_k \le 1$. Consequently, the first condition in (34) holds.

Since $\Gamma_k\Theta_k \le \frac{10c_p\beta n^\omega}{\mu b_k\mathbf{p}_k} \le \frac{5\beta c_p^2 n^{2\omega}}{\mu b_k}$, the second condition of (34) in Theorem 12 and (42) in Theorem 14 both hold if

$$\tfrac{5\beta c_p^2 n^{2\omega}}{\mu b} \le \Lambda.$$

Clearly, if we choose $b_k = b := \left\lfloor\frac{5\beta c_p^2 n^{2\omega}}{\mu\Lambda}\right\rfloor = \lfloor c_b n^{2\omega}\rfloor$ with $c_b := \frac{5\beta c_p^2}{\mu\Lambda}$, then this condition holds. Consequently, the second condition of (34) and (42) hold.

Since $B_K = 0$ and $E_0^2 = 0$, for a given $\epsilon > 0$, to guarantee $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$, from (35) we can impose that $\frac{2\Psi_0^2}{\mu^2(K+r-1)^2} \le \epsilon^2$. The last condition holds if we choose $K := \big\lfloor \frac{\sqrt{2}\Psi_0}{\mu\epsilon} \big\rfloor - 1$.

The expected total number of oracle calls is $\mathbb{E}\big[\mathcal{T}_K\big] := n + \mathbb{E}\big[\hat{\mathcal{T}}_K\big]$, where

$$\begin{aligned}
\mathbb{E}\big[\hat{\mathcal{T}}_K\big] &= \textstyle\sum_{k=0}^K (n\mathbf{p}_k + 2b)\\
&= n\textstyle\sum_{k=0}^K \mathbf{p}_k + 2b(K+1)\\
&= 2b(K+1) + n\textstyle\sum_{k=0}^{K_0} \mathbf{p}_k + n\sum_{k=K_0+1}^K \mathbf{p}_k\\
&\le 2b(K+1) + n\big[\textstyle\sum_{k=0}^{K_0} \frac{2}{c_p n^\omega} + 4\sum_{k=0}^{K_0} \frac{\mu}{\mu(k+r-1)-1} + \sum_{k=K_0+1}^K \frac{3}{c_p n^\omega}\big]\\
&\le \tfrac{2\sqrt{2}c_b\Psi_0 n^{2\omega}}{\mu\epsilon} + n\big[\tfrac{2(4c_p n^\omega - r + 1 + \mu^{-1})}{c_p n^\omega} + 4\ln(4c_p n^\omega - r + 2 + \mu^{-1}) + \tfrac{3\sqrt{2}\Psi_0}{c_p n^\omega \mu\epsilon}\big]\\
&\le 4n\big[2 + \ln(4c_p n^\omega)\big] + \tfrac{\sqrt{2}\Psi_0}{\mu\epsilon}\big(2c_b n^{2\omega} + \tfrac{3n^{1-\omega}}{c_p}\big).
\end{aligned}$$

Here, we have used $r - 1 - \frac{1}{\mu} \ge 1$ and $\sum_{k=0}^{K_0} \frac{\mu}{\mu(k+r-1)-1} \le \ln(1 + (r-1-\frac{1}{\mu})^{-1}K_0) \le \ln(K_0+1)$ in the last inequality. Therefore, we conclude that the expected total number of oracle calls F_i and $J_{\lambda T}$ evaluations in (VFOSA$_+$) is at most

$$\mathbb{E}\big[\mathcal{T}_K\big] := \Big\lfloor n + 4n\big[2 + \ln(4c_p n^\omega)\big] + \tfrac{\sqrt{2}\Psi_0}{\mu\epsilon}\big(2c_b n^{2\omega} + \tfrac{3n^{1-\omega}}{c_p}\big)\Big\rfloor$$

to achieve $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$. In particular, if we choose $\omega := \frac{1}{3}$, then we get $\mathbb{E}\big[\mathcal{T}_K\big] := \big\lfloor n + 4n\big[2 + \ln(4c_p n^{1/3})\big] + \frac{\sqrt{2}\Psi_0(3+2c_b c_p)n^{2/3}}{c_p\mu\epsilon}\big\rfloor$. ■

D.1.2 The proof of Corollary 19: The SAGA variant of VFOSA$_+$

Proof Since the SAGA estimator is used, by Lemma 6, we have

$$\kappa_k = \tfrac{\alpha b_k}{n}, \quad \Theta_k = \tfrac{(3-\alpha)n}{(1-\alpha)b_k^2} \ge \underline{\Theta} := \tfrac{3-\alpha}{(1-\alpha)n} > 0, \quad \text{and} \quad \sigma_k^2 = 0.$$

Note that B_K in Theorem 12 becomes $B_K := \frac{\Lambda}{\beta}\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} = 0$. Moreover, since $\hat{F}_i^0 = F_i x^0$ for all $i \in [n]$, we also have $\Delta_0 := \frac{1}{nb_0}\sum_{i=1}^n \|F_i x^0 - \hat{F}_i^0\|^2 \le \sigma_0^2 = 0$.

Next, let us choose $\alpha := \frac{1}{2}$ and $\Gamma_k := \frac{5\beta n^\omega}{\mu c_b} =: \underline{\Gamma} > 0$ for all $k \ge 0$ and given $\omega > 0$ and $c_b > 0$. We recall the first condition of (34) in Theorem 12 as follows:

$$\kappa_k = \tfrac{\alpha b_k}{n} = \tfrac{b_k}{2n} \ge 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_{k-1}(t_{k-1}-1)} + \tfrac{5\beta}{\mu\Gamma_k} = \tfrac{c_b}{n^\omega} + \tfrac{2\mu}{\mu(k+r-1)-1} - \tfrac{1+\mu}{(k+r-1)(\mu(k+r-1)-1)}.$$

Let us choose

$$b_k := \begin{cases} 2c_b n^{1-\omega} + \frac{4\mu n}{\mu(k+r-1)-1} & \text{if } \ k \le K_0 := \lfloor 4n^\omega + 1 + \mu^{-1} - r\rfloor,\\ 3c_b n^{1-\omega}, & \text{otherwise.}\end{cases}$$

Then, the last condition holds. Consequently, the first condition of (34) holds. Clearly, we have $b_k \le b_{k-1}$ for all $k \ge 1$. Moreover, using the update of b_k, we have

$$b_{k-1} - b_k = \tfrac{4\mu^2 n}{[\mu(k+r-1)-1][\mu(k+r-2)-1]} \le \tfrac{b_k b_{k-1}}{4n}.$$

Therefore, we finally get $b_{k-1} - \frac{(1-\alpha)b_k b_{k-1}}{2n} \leq b_k \leq b_{k-1}$, which verifies the condition of Lemma 6. Note that since $1 \leq b_k \leq n$ and $K_0 \geq 0$, we require $r > 5 + \frac{1}{\mu}$ and $n^\omega \geq \max\big\{\frac{\mu(r-1)-1}{4\mu}, \frac{2c_b[\mu(r-1)-1]}{\mu(r-5)-1}\big\}$ as stated in Corollary 19.

From the definition of Θ_k above and $b_k \geq 2c_b n^{1-\omega}$, we can easily show that $\Theta_k = \frac{(3-\alpha)n}{(1-\alpha)b_k^2} = \frac{5n}{b_k^2} \leq \frac{5}{4c_b^2 n^{1-2\omega}}$. Since $\Gamma_k\Theta_k \leq \frac{25\beta n^\omega}{4\mu c_b^2 n^{1-2\omega}}$, the second condition of (34) in Theorem 12 and the condition (42) in Theorem 14 both hold if $\frac{25\beta}{4\mu c_b^2 n^{1-3\omega}} \leq \Lambda$. Let us choose $\omega := \frac{1}{3}$ and $c_b := \frac{5}{2}\sqrt{\frac{\beta}{\mu\Lambda}}$. Then, the last condition holds. Consequently, the second condition of (34) and (42) are both satisfied.

Since $B_K = 0$ and $E_0^2 = 0$, for a given $\epsilon > 0$, to guarantee $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \leq \epsilon^2$, from (35), we can impose $\frac{2\Psi_0^2}{\mu^2(K+r-1)^2} \leq \epsilon^2$. This condition holds if we choose $K := \big\lfloor \frac{\sqrt{2}\Psi_0}{\mu\epsilon} \big\rfloor - 1$.

The expected total number of oracle calls becomes $\mathbb{E}\big[\mathcal{T}_K\big] := n + \mathbb{E}\big[\hat{\mathcal{T}}_K\big]$, where n is the number of F_i evaluations for the first epoch, and

$$\begin{aligned}\mathbb{E}\big[\hat{\mathcal{T}}_K\big] &= \textstyle\sum_{k=0}^K b_k = \sum_{k=0}^{K_0} b_k + \sum_{k=K_0+1}^K b_k\\ &\leq 2c_b n^{2/3}(K_0+1) + \textstyle\sum_{k=0}^{K_0}\frac{4\mu n}{\mu(k+r-1)-1} + 3c_b(K-K_0)n^{2/3}\\ &\leq 8c_b n^{2/3}n^{1/3} + 4n\ln(K_0+1) + 3c_b K n^{2/3}\\ &\leq 8c_b n + 4n\ln(4n^{1/3}) + \frac{3\sqrt{2}c_b\Psi_0 n^{2/3}}{\mu\epsilon}.\end{aligned}$$

Thus, we get $\mathbb{E}\big[\mathcal{T}_K\big] := \big\lfloor [8c_b + 4\ln(4n^{1/3})]n + \frac{3\sqrt{2}c_b\Psi_0 n^{2/3}}{\mu\epsilon}\big\rfloor$. We conclude that the expected total number of F_i and $J_{\lambda T}$ evaluations of (VFOSA$_+$) is at most $\mathbb{E}\big[\mathcal{T}_K\big] := \mathcal{O}\big(n\ln(n) + \frac{n^{2/3}}{\epsilon}\big)$ to achieve $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \leq \epsilon^2$. ■

D.1.3 The proof of Corollary 20: The L-SARAH variant of VFOSA$_+$

Proof Since the L-SARAH estimator is used and $\bar{F}x^k = Fx^k$, by Lemma 7, we have

$$\kappa_k = \mathbf{p}_k, \quad \Theta_k = \tfrac{1}{b_k} \geq \underline{\Theta} := \tfrac{1}{n} > 0, \quad \text{and} \quad \sigma_k^2 = 0.$$

The quantity B_K in Theorem 12 becomes $B_K := \frac{\Lambda}{\beta}\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} = 0$. Moreover, since $\widetilde{F}^0 := Fx^0$, we have $\|\widetilde{F}^0 - Fx^0\|^2 \leq \sigma_0^2 = 0$.

Next, let us choose $\Gamma_k = \frac{5c_p\beta n^\omega}{\mu} =: \underline{\Gamma} > 0$ for all $k \geq 0$ and given $c_p > 0$ and $\omega > 0$. We recall the first condition of (34) in Theorem 12 as follows:

$$\kappa_k = \mathbf{p}_k \geq 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_{k-1}(t_{k-1}-1)} + \tfrac{5\beta}{\mu\Gamma_k} = \tfrac{1}{c_p n^\omega} + \tfrac{2\mu}{\mu(k+r-1)-1} - \tfrac{1+\mu}{(k+r-1)(\mu(k+r-1)-1)}.$$

If $r > 3 + \frac{1}{\mu}$ and $n^\omega \geq \frac{1}{c_p}\max\{\frac{\mu(r-1)-1}{\mu(r-3)-1}, \frac{\mu(r-1)-1}{2\mu}\}$, then by choosing

$$\mathbf{p}_k := \begin{cases} \frac{1}{c_p n^\omega} + \frac{2\mu}{\mu(k+r-1)-1} & \text{if } k \leq K_0 := \lfloor 2c_p n^\omega - r + 1 + \mu^{-1}\rfloor,\\ \frac{2}{c_p n^\omega} & \text{otherwise,}\end{cases}$$

we have $0 < \mathbf{p}_k \leq 1$ and the last condition holds. Thus the first condition of (34) holds.

Since $\Gamma_k\Theta_k = \frac{5\beta c_p n^\omega}{\mu b_k}$, the second condition of (34) in Theorem 12 and the condition (42) in Theorem 14 both hold if $\frac{5c_p\beta n^\omega}{\mu b_k} \le \Lambda$. Clearly, if we choose $b_k = b := \lfloor \frac{5\beta c_p n^\omega}{\mu\Lambda} \rfloor = \lfloor c_b n^\omega \rfloor$ with $c_b := \frac{5\beta c_p}{\mu\Lambda}$, then this condition holds. Consequently, the second condition of (34) and (42) are both satisfied.

Since $B_K = 0$ and $E_0^2 = 0$, for a given $\epsilon > 0$, to guarantee $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$, from (35), we can impose $\frac{2\Psi_0^2}{\mu^2(K+r-1)^2} \le \epsilon^2$. This condition holds if we choose $K := \lfloor \frac{\sqrt{2}\Psi_0}{\mu\epsilon} \rfloor - 1$.

We can estimate the expected total number of oracle calls as $\mathbb{E}\big[\mathcal{T}_K\big] := n + \mathbb{E}\big[\hat{\mathcal{T}}_K\big]$, where

$$\begin{aligned}
\mathbb{E}\big[\hat{\mathcal{T}}_K\big] &= \textstyle\sum_{k=0}^K [\mathbf{p}_k n + 2(1-\mathbf{p}_k)b] \\
&= n\textstyle\sum_{k=0}^K \mathbf{p}_k + 2b\sum_{k=0}^K (1-\mathbf{p}_k) \\
&= 2b(K+1) + (n-2b)\textstyle\sum_{k=0}^{K_0}\mathbf{p}_k + (n-2b)\sum_{k=K_0+1}^K \mathbf{p}_k \\
&\le 2b(K+1) + n\big[\textstyle\sum_{k=0}^{K_0}\frac{1}{c_p n^\omega} + 2\sum_{k=0}^{K_0}\frac{\mu}{\mu(k+r-1)-1} + \sum_{k=K_0+1}^K \frac{2}{c_p n^\omega}\big] \\
&\le \tfrac{\sqrt{2}c_b\Psi_0 n^\omega}{\mu\epsilon} + n\big[\tfrac{2c_p n^\omega - r + 1 + \mu^{-1}}{c_p n^\omega} + 2\ln(2c_p n^\omega - r + 1 + \mu^{-1}) + \tfrac{2\sqrt{2}\Psi_0}{c_p n^\omega \mu\epsilon}\big] \\
&\le 2n\big[2 + \ln(c_p n^\omega)\big] + \tfrac{\sqrt{2}\Psi_0}{\mu\epsilon}\big(c_b n^\omega + \tfrac{2n^{1-\omega}}{c_p}\big).
\end{aligned}$$

Here, we have used $r - 1 - \frac{1}{n} \ge 1$ in the last inequality. Hence, we conclude that the expected total number of oracle calls F_i and $J_{\lambda T}$ evaluations of (VFOSA$_+$) is at most

$$\mathbb{E}\big[\mathcal{T}_K\big] := \Big\lfloor n + 2n\big[2 + \ln(c_p n^\omega)\big] + \tfrac{\sqrt{2}\Psi_0}{\mu\epsilon}\big(c_b n^\omega + \tfrac{2n^{1-\omega}}{c_p}\big)\Big\rfloor$$

to achieve $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$. Clearly, if we choose $\omega = \frac{1}{2}$, then we obtain $\mathbb{E}\big[\mathcal{T}_K\big] := \mathcal{O}\big(n\ln(n^{1/2}) + \frac{n^{1/2}}{\epsilon}\big)$. ■

D.2 The expectation setting (E)

This appendix presents the full proof of Corollaries 22, 23, and 24, respectively.

D.2.1 The proof of Corollary 22: The L-SVRG variant of VFOSA$_+$

Proof Since L-SVRG is used, substituting $\tau := 1$ and $\alpha := \frac{1}{2}$ into Lemma 5, we have

$$\Delta_k := 2\hat{\Delta}_k + \tfrac{2\sigma^2}{n_k}, \quad \kappa_k := \alpha\mathbf{p}_k = \tfrac{\mathbf{p}_k}{2}, \quad \Theta_k := \tfrac{2}{(1-\alpha)b_k\mathbf{p}_k} = \tfrac{4}{b_k\mathbf{p}_k}, \quad \text{and} \quad \sigma_k^2 := \tfrac{2\alpha\mathbf{p}_k\sigma^2}{n_k} = \tfrac{\mathbf{p}_k\sigma^2}{n_k},$$

to fulfill the $\mathbf{VR}(\kappa_k, \Theta_k, \Delta_k, \sigma_k)$ condition of Definition 4.

Next, let us choose $\Gamma_k := \frac{5\beta}{\mu\epsilon^\omega} =: \underline{\Gamma} > 0$ for a given tolerance $\epsilon \in (0,1)$ and a given $\omega > 0$. We recall the first condition of (34) in Theorem 12 as follows:

$$\kappa_k = \alpha\mathbf{p}_k = \tfrac{\mathbf{p}_k}{2} \ge 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_{k-1}(t_{k-1}-1)} + \tfrac{5\beta}{\mu\Gamma_k} = \epsilon^\omega + \tfrac{2\mu}{\mu(k+r-1)-1} - \tfrac{1+\mu}{(k+r-1)(\mu(k+r-1)-1)}.$$

From this relation, we can see that if we choose

$$\mathbf{p}_k := 2\epsilon^\omega + \tfrac{4\mu}{\mu(k+r-1)-1} \ge \underline{\mathbf{p}} := 2\epsilon^\omega,$$

then the first condition in (34) holds and $\mathbf{p}_k \le 1$, provided that $r > 5 + \frac{1}{\mu}$ and $\epsilon^\omega \le \frac{\mu(r-5)-1}{2\mu(r-1)-2}$.

Since $\Gamma_k\Theta_k \le \frac{10\beta}{\mu(1-\alpha)b\mathbf{p}_k\epsilon^\omega} \le \frac{10\beta}{\mu b\epsilon^{2\omega}}$, the second condition of (34) in Theorem 12 and the condition (42) in Theorem 14 both hold if $\frac{10\beta}{\mu b\epsilon^{2\omega}} \le \Lambda$. Therefore, if we choose $b_k = b := \lfloor \frac{10\beta}{\mu\Lambda\epsilon^{2\omega}} \rfloor = \lfloor \frac{c_b}{\epsilon^{2\omega}} \rfloor$ with $c_b := \frac{10\beta}{\mu\Lambda}$, then the last condition holds. Consequently, the second condition of (34) and (42) are both satisfied. Moreover, since $b_k = b > 0$ for all $k \ge 0$, we have $\Theta_k \ge \frac{4}{b} =: \underline{\Theta} > 0$, which fulfills (42).

If we fix $n_k = n > 0$, then the quantity B_K in Theorem 12 becomes

$$\begin{aligned} B_K &:= \tfrac{\Lambda}{\beta}\textstyle\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} = \frac{5\sigma^2}{\mu n\epsilon^{2\omega}}\sum_{k=0}^K \mathbf{p}_k^2 t_{k-1}(t_{k-1}-1)\\ &\le \tfrac{10\sigma^2}{n\epsilon^\omega}\textstyle\sum_{k=0}^K \big(\epsilon^\omega + \frac{2\mu}{\mu(k+r-1)-1}\big)(k+r-1)[\mu(k+r-1)-1]\\ &\le \tfrac{10\mu\sigma^2}{n\epsilon^\omega}\textstyle\sum_{k=0}^K \big[\epsilon^\omega(k+r-1)^2 + 2(k+r-1)\big]\\ &\le \tfrac{10\mu\sigma^2}{n}\big(K+\epsilon^{-\omega}\big)(K+r-1)^2. \end{aligned}$$

Moreover, from the construction (L-SVRG) of $\widetilde{F}^0$, we have $E_0^2 = \frac{\mu r^2(\mu\Gamma_0+\beta)\sigma^2}{2\beta n}$. We also have $\Gamma_0\mu + \beta = \frac{5\beta}{\epsilon^\omega} + \beta \le \frac{6\beta}{\epsilon^\omega}$. Using these bounds, to guarantee $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$, from (35), we can impose

$$\begin{aligned} \mathbb{E}\big[\|G_\lambda x^K\|^2\big] &\le \tfrac{2(\Psi_0^2+B_{K-1})}{\mu^2(K+r-1)^2} + \tfrac{\mu r^2(\Gamma_0\mu+\beta)}{\mu^2 n\beta(K+r-1)^2}\sigma^2\\ &\le \tfrac{2\Psi_0^2}{\mu^2(K+r-1)^2} + \tfrac{6r^2\sigma^2}{\mu(K+r-1)^2\epsilon^\omega n} + \tfrac{3\sigma^2}{\mu n}\big(K+\epsilon^{-\omega}\big)\\ &\le \epsilon^2. \end{aligned}$$

This condition holds if

$$\tfrac{2\Psi_0^2}{\mu^2(K+r-1)^2} \le \tfrac{\epsilon^2}{4}, \quad \tfrac{3\sigma^2 K}{\mu n} \le \tfrac{\epsilon^2}{4}, \quad \tfrac{3\sigma^2}{n\epsilon^\omega} \le \tfrac{\epsilon^2}{4}, \quad \text{and} \quad \tfrac{6r^2\sigma^2}{\mu(K+r-1)^2\epsilon^\omega n} \le \tfrac{\epsilon^2}{4}.$$

These four conditions are simultaneously satisfied if we choose

$$K := \big\lfloor \tfrac{2\sqrt{2}\Psi_0}{\mu\epsilon} - 1\big\rfloor \quad \text{and} \quad n \ge \sigma^2\max\Big\{\tfrac{3\mu r^2}{\Psi_0^2\epsilon^\omega}, \tfrac{12}{\epsilon^{2+\omega}}, \tfrac{24\sqrt{2}\Psi_0}{\mu^2\epsilon^3}\Big\}.$$

Hence, we can choose $\omega := 1$ and $n := \frac{c_n}{\epsilon^3} = \mathcal{O}\big(\frac{1}{\epsilon^3}\big)$ for given $c_n := 12\sigma^2\max\{1, \frac{2\sqrt{2}\Psi_0}{\mu^2}\}$.

Since we fix $n_k = n > 0$ and $b_k = b > 0$ for all $k \ge 0$, the expected total number of oracle calls becomes $\mathbb{E}\big[\mathcal{T}_K\big] = n + \mathbb{E}\big[\hat{\mathcal{T}}_K\big]$, where

$$\begin{aligned} \mathbb{E}\big[\hat{\mathcal{T}}_K\big] &= \textstyle\sum_{k=0}^K(n\mathbf{p}_k + 2b) = 2b(K+1) + n\sum_{k=0}^K \mathbf{p}_k\\ &\le 2b(K+1) + 2n\big[\textstyle\sum_{k=0}^K \epsilon + 2\sum_{k=0}^K \frac{\mu}{\mu(k+r-1)-1}\big]\\ &\le \tfrac{4c_b\sqrt{2}\Psi_0}{\mu\epsilon^3} + \tfrac{2c_n}{\epsilon^3}\big[\epsilon(K+1) + 2\ln(K+1)\big]\\ &\le \tfrac{4\sqrt{2}c_b\Psi_0}{\mu\epsilon^3} + \tfrac{2c_n}{\epsilon^3}\big[\tfrac{2\sqrt{2}\Psi_0}{\mu} + 2\ln\big(\tfrac{2\sqrt{2}\Psi_0}{\mu\epsilon}\big)\big]. \end{aligned}$$

Therefore, we obtain $\mathbb{E}\big[\hat{\mathcal{T}}_K\big] = \mathcal{O}\big(\epsilon^{-3} + \epsilon^{-3}\ln(\epsilon^{-1})\big) = \widetilde{\mathcal{O}}(\epsilon^{-3})$. This inequality shows that the expected total number of oracle calls $\mathbf{F}(\cdot,\xi)$ and $J_{\lambda T}$ evaluations of (VFOSA$_+$) is at most $\mathbb{E}\big[\mathcal{T}_K\big] := \mathcal{O}\big(\epsilon^{-3} + \epsilon^{-3}\ln(\epsilon^{-1})\big)$ to achieve $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$. ■

D.2.2 The proof of Corollary 23: The L-SARAH variant of VFOSA$_+$

Proof Since the L-SARAH estimator is used, by Lemma 7, we have

$$\kappa_k := \mathbf{p}_k, \quad \Theta_k := \tfrac{1}{b_k}, \quad \text{and} \quad \sigma_k^2 := \tfrac{\mathbf{p}_k\sigma^2}{n}.$$

Next, let us choose $\Gamma_k = \frac{5\beta}{\mu\epsilon^\omega} =: \underline{\Gamma} > 0$ for all $k \geq 0$, a given tolerance $\epsilon \in (0,1)$, and $\omega > 0$. We recall the first condition of (34) in Theorem 12 as follows:

$$\begin{aligned}\kappa_k = \mathbf{p}_k &\geq 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_k(t_k-1)} + \tfrac{5\beta}{\mu\Gamma_k} = \tfrac{5\beta}{\mu\Gamma_k} + \tfrac{2\mu}{\mu(k+r-1)-1} - \tfrac{1+\mu}{(k+r-1)(\mu(k+r-1)-1)}\\ &= \epsilon^\omega + \tfrac{2\mu}{\mu(k+r-1)-1} - \tfrac{1+\mu}{(k+r-1)(\mu(k+r-1)-1)}.\end{aligned}$$

This condition holds if we choose $\mathbf{p}_{k+1} := \epsilon^\omega + \frac{2\mu}{\mu(k+r-1)-1}$, provided that $r > 3 + \frac{1}{\mu}$ and $0 < \epsilon^\omega \leq \frac{\mu(r-3)-1}{\mu(r-1)-1}$. In addition, we also have $0 < \underline{\mathbf{p}} := \epsilon^\omega \leq \mathbf{p}_k \leq 1$.

Since $\Gamma_k\Theta_k = \frac{5\beta}{\mu b_k\epsilon^\omega}$, the second condition of (34) in Theorem 12 and the condition (42) in Theorem 14 both hold if $\frac{5\beta}{\mu b_k\epsilon^\omega} \leq \Lambda$. Therefore, if we choose $b_k = b := \big\lfloor \frac{5\beta}{\mu\Lambda\epsilon^\omega}\big\rfloor = \big\lfloor \frac{c_b}{\epsilon^\omega}\big\rfloor$ with $c_b := \frac{5\beta}{\mu\Lambda}$, then the last condition holds. Consequently, the second condition of (34) and (42) are both satisfied. Since $b_k = b > 0$, we have $\Theta_k \geq \frac{1}{b} =: \underline{\Theta} > 0$.

If we fix $n_k = n > 0$, then the quantity B_K in Theorem 12 becomes

$$\begin{aligned}B_K &:= \tfrac{\Lambda}{\beta}\textstyle\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k} = \frac{5\sigma^2}{2\mu n\epsilon^\omega}\sum_{k=0}^n \mathbf{p}_k t_{k-1}(t_{k-1}-1)\\ &= \tfrac{5\sigma^2}{2n\epsilon^\omega}\textstyle\sum_{k=0}^K \big(\epsilon^\omega + \frac{2\mu}{\mu(k+r-1)-1}\big)(k+r-1)[\mu(k+r-1)-1]\\ &\leq \tfrac{5\mu\sigma^2}{2n}\big(K + 2\epsilon^{-\omega}\big)(K+r-1)^2.\end{aligned}$$

Moreover, we also have $E_0^2 = \frac{\mu r^2(\mu\Gamma_0+\beta)\sigma^2}{2\beta n}$ and $\Gamma_0\mu + \beta = \frac{5\beta}{\epsilon^\omega} + \beta \leq \frac{6\beta}{\epsilon^\omega}$. Using these bounds, from (35), to guarantee $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \leq \epsilon^2$, we impose

$$\begin{aligned}\mathbb{E}\big[\|G_\lambda x^K\|^2\big] &\leq \tfrac{2(\Psi_0^2 + B_{K-1})}{\mu^2(K+r-1)^2} + \tfrac{\mu r^2(\Gamma_0\mu+\beta)}{\mu^2 n\beta(K+r-1)^2}\sigma^2\\ &\leq \tfrac{2\Psi_0^2}{\mu^2(K+r-1)^2} + \tfrac{3\sigma^2}{\mu n}\big(K + 2\epsilon^{-\omega}\big) + \tfrac{6r^2\sigma^2}{\mu(K+r-1)^2\epsilon^\omega n}\\ &\leq \epsilon^2.\end{aligned}$$

This condition holds if

$$\tfrac{2\Psi_0^2}{\mu^2(K+r-1)^2} \leq \tfrac{\epsilon^2}{4}, \quad \tfrac{3\sigma^2 K}{\mu n} \leq \tfrac{\epsilon^2}{4}, \quad \tfrac{6\sigma^2}{\mu\beta n\epsilon^\omega} \leq \tfrac{\epsilon^2}{4}, \quad \text{and} \quad \tfrac{6r^2\sigma^2}{\mu(K+r-1)^2\epsilon^\omega n} \leq \tfrac{\epsilon^2}{4}.$$

These four conditions are simultaneously satisfied if we choose

$$K := \big\lfloor \tfrac{2\sqrt{2}\Psi_0}{\mu\epsilon} - 1\big\rfloor \quad \text{and} \quad n \geq \sigma^2\max\Big\{\tfrac{24\sqrt{2}\Psi_0}{\mu^2\epsilon^3}, \tfrac{24}{\mu\beta\epsilon^{2+\omega}}, \tfrac{3r^2\mu}{\Psi_0^2\epsilon^\omega}\Big\}.$$

Thus, we can choose $\omega = 1$ and $n := \frac{c_n}{\epsilon^3} = \mathcal{O}\big(\frac{1}{\epsilon^3}\big)$ for given $c_n := 24\sigma^2\max\Big\{\frac{\sqrt{2}\Psi_0}{\mu^2}, \frac{1}{\mu\beta}\Big\}$.

Since we fix $n_k = n > 0$ and $b_k = b > 0$, the expected total number of oracle calls becomes $\mathbb{E}\big[\mathcal{T}_K\big] = n + \mathbb{E}\big[\hat{\mathcal{T}}_K\big]$, where

$$\begin{aligned}\mathbb{E}\big[\hat{\mathcal{T}}_K\big] &= \textstyle\sum_{k=0}^K[n\mathbf{p}_k + 2b(1-\mathbf{p}_k)] = n\sum_{k=0}^K\mathbf{p}_k + 2b\sum_{k=0}^K(1-\mathbf{p}_k)\\ &= 2b(K+1) + (n-2b)\textstyle\sum_{k=0}^K\mathbf{p}_k\\ &\le 2b(K+1) + n\big[\textstyle\sum_{k=0}^K\epsilon + 2\sum_{k=0}^K\frac{\mu}{\mu(k+r-1)-1}\big]\\ &\le \tfrac{4c_b\sqrt{2}\Psi_0}{\mu\epsilon^2} + \tfrac{c_n}{\epsilon^3}\big[\epsilon(K+1) + 2\ln(K+r)\big]\\ &\le \tfrac{4c_b\sqrt{2}\Psi_0}{\mu\epsilon^2} + \tfrac{c_n}{\epsilon^3}\big[\tfrac{2\sqrt{2}\Psi_0}{\mu} + 2\ln\big(\tfrac{3\sqrt{2}\Psi_0}{\mu\epsilon}\big)\big].\end{aligned}$$

Therefore, we obtain $\mathbb{E}\big[\hat{\mathcal{T}}_K\big] = \mathcal{O}\big(\epsilon^{-2} + \epsilon^{-3} + \epsilon^{-3}\ln(\epsilon^{-1})\big) = \widetilde{\mathcal{O}}(\epsilon^{-3})$. This inequality shows that the expected total number of oracle calls $\mathbf{F}(\cdot,\xi)$ and $J_{\lambda T}$ evaluations of (VFOSA$_+$) is at most $\mathbb{E}\big[\mathcal{T}_K\big] := \mathcal{O}\big(\epsilon^{-2} + \epsilon^{-3} + \epsilon^{-3}\ln(\epsilon^{-1})\big)$ to achieve $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \le \epsilon^2$. ■

D.2.3 The proof of Corollary 24: The HSGD variant of VFOSA$_+$

Proof Since (HSGD) is used for Fx^k, by Lemma 8, we have

$$\kappa_k := 1 - (1-\tau_k)^2, \quad \Theta_k := \tfrac{2(1-\tau_k)^2}{b_k}, \quad \text{and} \quad \sigma_k^2 := \tfrac{2\tau_k^2\sigma^2}{\hat{b}_k}.$$

Now, let us choose $\Gamma_k = \frac{5\beta t_k(t_k-1)}{\theta\mu t_{k-1}(t_{k-1}-1)} \ge \frac{5\beta}{\theta\mu} =: \underline{\Gamma} > 0$ for all $k \ge 0$ and some $\theta \in (0,1)$. Then, the first condition of (34) in Theorem 12 is equivalent to

$$\kappa_k = 1 - (1-\tau_k)^2 \ge 1 - \tfrac{\Gamma_{k-1}t_{k-2}(t_{k-2}-1)}{\Gamma_k t_{k-1}(t_{k-1}-1)} + \tfrac{5\beta}{\mu\Gamma_k} = 1 - \tfrac{(1-\theta)t_{k-1}(t_{k-1}-1)}{t_k(t_k-1)}.$$

This condition is equivalent to

$$\tfrac{(1-\theta)t_{k-1}(t_{k-1}-1)}{t_k(t_k-1)} \ge (1-\tau_k)^2.$$

If we choose $\tau_k := 1 - \sqrt{\frac{(1-\theta)t_{k-1}(t_{k-1}-1)}{t_k(t_k-1)}}$ as stated in Corollary 24, then the last condition holds. Consequently, the first condition of (34) in Theorem 12 holds.

Since $\Theta_k = \frac{2(1-\tau_k)^2}{b_k} = \frac{2(1-\theta)t_{k-1}(t_{k-1}-1)}{b_k t_k(t_k-1)}$, we have $\Gamma_k\Theta_k = \frac{10\beta(1-\theta)}{\mu\theta b_k}$. The second condition of (34) in Theorem 12 and the condition (42) in Theorem 14 both hold if $\frac{10\beta(1-\theta)}{\mu\theta b_k} \le \Lambda$. Therefore, if we choose $b_k = b \ge \frac{10\beta}{\mu\Lambda\theta}$, then the last condition holds. Consequently, the second condition of (34) and (42) are both satisfied. Moreover, since $\tau_k \ge 1 - \sqrt{1-\theta}$, $b_k = b > 0$ and $\hat{b}_k = \hat{b} > 0$ for all $k \ge 0$, we have $\Theta_k \ge \underline{\Theta} := \frac{\hat{c}(1-\theta)}{b} > 0$ for a given $\hat{c} > 0$.

Now, since $t_k = t_{k-1} + \mu$, if $t_{k-1} \ge \frac{1}{1-\mu^2}$ (which holds if $r \ge 1 + \frac{1}{\mu(1-\mu^2)}$), then we have $\sqrt{t_k(t_k-1)} \le \sqrt{t_{k-1}(t_{k-1}-1)} + 1$. Therefore, we get

$$\begin{aligned}\mathcal{T}_{[1]} &:= \big[\sqrt{t_k(t_k-1)} - \sqrt{(1-\theta)t_{k-1}(t_{k-1}-1)}\big]^2\\ &\le \big[(1-\sqrt{1-\theta})\sqrt{t_{k-1}(t_{k-1}-1)} + 1\big]^2\\ &\le \big(\theta t_{k-1} + 1\big)^2 = \mu^2\theta^2(k+r-1)^2 + 2\mu\theta(k+r-1) + 1.\end{aligned}$$

Since we choose $r \geq 5 + \frac{1}{\mu}$, we also have $r \geq 1 + \frac{1}{\mu(1-\mu^2)}$ and $\Gamma_0 \leq \frac{5c_0\beta}{\mu\theta}$ for a constant $c_0 > 0$.

In this case, we can bound the quantity B_K in Theorem 12 as follows:

$$\begin{aligned}
B_K &:= \tfrac{\Lambda}{\beta}\textstyle\sum_{k=0}^K t_{k-1}(t_{k-1}-1)\frac{\sigma_k^2}{\Theta_k}\\
&= \tfrac{\Lambda\sigma^2 b}{\beta(1-\theta)\hat{b}}\textstyle\sum_{k=0}^K t_k(t_k-1)\Big(1 - \sqrt{\frac{(1-\theta)t_{k-1}(t_{k-1}-1)}{t_k(t_k-1)}}\Big)^2\\
&= \tfrac{\Lambda\sigma^2 b}{\beta(1-\theta)\hat{b}}\textstyle\sum_{k=0}^K \big[\sqrt{t_k(t_k-1)} - \sqrt{(1-\theta)t_{k-1}(t_{k-1}-1)}\big]^2\\
&\leq \tfrac{\Lambda\sigma^2 b}{\beta(1-\theta)\hat{b}}\textstyle\sum_{k=0}^K \big[\mu^2\theta^2(k+r-1)^2 + 2\mu\theta(k+r-1) + 1\big]\\
&\leq \tfrac{\Lambda\sigma^2 b}{\beta(1-\theta)\hat{b}}\big[\mu^2\theta^2 K(K+r-1)^2 + 2\mu\theta(K+r-1)^2 + (K+1)\big].
\end{aligned}$$

In addition, we also have $E_0^2 = \frac{\mu r^2(\mu\Gamma_0+\beta)\sigma^2}{2\beta n_0}$. Using these bounds and $\Gamma_0 \leq \frac{5c_0\beta}{\mu\theta}$, to guarantee $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \leq \epsilon^2$ for any $\epsilon \in (0, 1/2]$, from (35), we impose

$$\begin{aligned}
\mathbb{E}\big[\|G_\lambda x^K\|^2\big] &\leq \tfrac{2(\Psi_0^2 + B_{K-1})}{\mu^2(K+r-1)^2} + \tfrac{\mu r^2(\Gamma_0\mu+\beta)}{\mu^2 n_0\beta(K+r-1)^2}\sigma^2\\
&\leq \tfrac{2\Psi_0^2}{\mu^2(K+r-1)^2} + \tfrac{2\Lambda\sigma^2 b}{\mu^2\beta(1-\theta)\hat{b}}\big(\mu^2\theta^2 K + 2\mu\theta + \tfrac{1}{K}\big) + \tfrac{r^2\sigma^2(5c_0+\theta)}{(K+r-1)^2\mu\theta n_0}\\
&\leq \epsilon^2.
\end{aligned}$$

If we assume that $\theta \in (0, 1/2]$ and $\theta \leq c_0$, then the last condition holds if

$$\tfrac{2\Psi_0^2}{\mu^2(K+r-1)^2} \leq \tfrac{\epsilon^2}{5}, \quad \tfrac{4\Lambda\sigma^2\theta^2 bK}{\beta\hat{b}} \leq \tfrac{\epsilon^2}{5}, \quad \tfrac{8\Lambda\theta\sigma^2 b}{\mu\beta\hat{b}} \leq \tfrac{\epsilon^2}{5}, \quad \tfrac{4\Lambda\sigma^2 b}{\mu^2\beta\hat{b}K} \leq \tfrac{\epsilon^2}{5}, \quad \text{and} \quad \tfrac{6c_0 r^2\sigma^2}{(K+r-1)^2\mu\theta n_0} \leq \tfrac{\epsilon^2}{5}.$$

These five conditions are simultaneously satisfied if we choose

$$K := \big\lfloor \tfrac{\sqrt{10}\Psi_0}{\mu\epsilon} - 1\big\rfloor, \quad \theta := \epsilon, \quad n_0 := \tfrac{3c_0\mu r^2\sigma^2}{\Psi_0^2\epsilon}, \quad \text{and} \quad \tfrac{\hat{b}}{b} \geq \tfrac{2\sqrt{10}\Lambda\sigma^2}{\mu\beta\epsilon}\max\big\{10\Psi_0, 2\sqrt{10}, \tfrac{1}{\Psi_0}\big\}.$$

Combining this condition and $b_k = b \geq \frac{3\beta}{\mu\Lambda\theta} = \frac{3\beta}{\mu\Lambda\epsilon}$ above, we conclude that $b_k = b := \big\lfloor \frac{c_b}{\epsilon}\big\rfloor$ for $c_b := \frac{3\beta}{\mu\Lambda}$ and $\hat{b}_k = \hat{b} := \lfloor \frac{\hat{c}_b}{\epsilon^2}\rfloor$ for $\hat{c}_b := \frac{6\sqrt{10}\sigma^2}{\mu^2}\max\Big\{10\Psi_0, 2\sqrt{10}, \frac{1}{\Psi_0}\Big\}$ and all $k \geq 1$, and $n_0 := \big\lfloor \frac{c_n}{\epsilon}\big\rfloor$ for $c_n := \frac{3c_0\mu r^2\sigma^2}{\Psi_0^2}$. Since $\epsilon \in (0, 1/2]$ and $\theta = \epsilon$, we also have $\theta \in (0, 1/2]$.

Finally, since $b_k = b > 0$ and $\hat{b}_k = \hat{b} > 0$ for all $k \geq 0$, the expected total number of oracle calls is $\mathbb{E}\big[\mathcal{T}_K\big] = \mathbb{E}\big[\mathcal{T}_0\big] + \mathbb{E}\big[\hat{\mathcal{T}}_k\big]$, where $\hat{\mathcal{T}}_k$ is

$$\mathbb{E}\big[\hat{\mathcal{T}}_K\big] = (2b + \hat{b})(K+1) = \big(\tfrac{2c_b}{\epsilon} + \tfrac{\hat{c}_b}{\epsilon^2}\big)\tfrac{\sqrt{10}\Psi_0}{\mu\epsilon} \leq \tfrac{(2c_b+\hat{c}_b)\sqrt{10}\Psi_0}{\mu\epsilon^3},$$

and $\mathbb{E}\big[\mathcal{T}_0\big] = 2b + n_0 = \mathcal{O}\big(\epsilon^{-1}\big)$ is the expected total number of oracle calls for evaluating $\widetilde{F}^0$. Overall, the expected total number of oracle calls $\mathbf{F}(\cdot, \xi)$ and $J_{\lambda T}$ evaluations of (VFOSA$_+$) is at most $\mathbb{E}\big[\mathcal{T}_K\big] := \mathcal{O}\big(\epsilon^{-1} + \epsilon^{-3}\big)$ to achieve $\mathbb{E}\big[\|G_\lambda x^K\|^2\big] \leq \epsilon^2$. ■

Appendix E. The Convergence Analysis of Algorithm 2

This appendix provides the full proofs of the results in Section 5.

E.1 The proof of Theorem 25: Key bounds

Proof To analyze the convergence of (VFOSA_), we construct the following functions:

$$\begin{aligned}\widehat{\mathcal{L}}_k &:= \beta a_k\|S_\lambda u^k\|^2 + t_{k-1}\langle S_\lambda u^k, u^k - s^k\rangle + \tfrac{c_k}{2\nu\beta}\|s^k - u^\star\|^2,\\ \widehat{\mathcal{Q}}_k &:= \widehat{\mathcal{L}}_k + \big[\bar{\beta} - (1+s)\beta\big]t_{k-1}(t_{k-1}-1)\|S_\lambda u^k - S_\lambda u^{k-1}\|^2\\ &\quad + \Lambda L t_{k-1}(t_{k-1}-1)\langle Fx^k - Fx^{k-1}, x^k - x^{k-1}\rangle,\\ \widehat{\mathcal{P}}_k &:= \widehat{\mathcal{Q}}_k + \tfrac{[\mu(1-\kappa_k)\Gamma_k+\beta]t_{k-1}(t_{k-1}-1)}{2\mu}\Delta_{k-1},\end{aligned}$$

where $u^\star \in \mathrm{zer}(S_\lambda)$ arbitrary, $a_k := t_{k-1}[t_{k-1} - 1 - s(1-\nu)]$, $\mu \in (0,1]$, $s > 0$, c_k is given in (26), and $\Gamma_k > 0$ are given parameters. The quantity Δ_k and the parameter κ_k are given in Definition 4. The functions $\widehat{\mathcal{L}}_k$, $\widehat{\mathcal{Q}}_k$, and $\widehat{\mathcal{P}}_k$ are similar to the ones $\mathcal{L}_k$, $\mathcal{Q}_k$, and $\mathcal{P}_k$ in (25), respectively.

In this case, $S_\lambda u$ and (u^k, s^k) play the same role as $G_\lambda x$ and (x^k, z^k) in Section 4 and the results of Theorem 12 still hold for $S_\lambda u$ and $\{(u^k, s^k)\}$. Using the relation $\|Fx^k + \xi^k\| = \|F(J_{\lambda T}u^k) + \lambda^{-1}(u^k - J_{\lambda T}u^k)\| = \|S_\lambda u^k\|$, we obtain (46) from (35). The estimates in (47) follow from (37) and the relation $\|Fx^k + \xi^k\| = \|S_\lambda u^k\|$. Finally, since $u \in J_{\lambda T}u + \lambda T(J_{\lambda T}u)$, it is obvious to check that $\xi^k := \frac{1}{\lambda}(u^k - J_{\lambda T}u^k) = \frac{1}{\lambda}(u^k - x^k) \in T(J_{\lambda T}u^k) = Tx^k$. ■

E.2 The proof of Theorem 26: The $\mathcal{O}(1/k^2)$ and $o(1/k^2)$-convergence rates

Proof The conclusions of Theorem 26 follow from the results of Theorem 13, respectively and the relations $\|x^{k+1} - x^k\| = \|J_{\lambda T}u^{k+1} - J_{\lambda T}u^k\| \le \|u^{k+1} - u^k\|$ and $\|Fx^k + \xi^k\| = \|S_\lambda u^k\|$ for $\xi^k := \frac{1}{\lambda}(u^k - x^k) \in Tx^k$. We omit the details to avoid a repetition. ■

E.3 The proof of Theorem 27: Almost sure convergence

Proof First, the $o(1/k^2)$ almost sure convergence rates also follow from those in Theorem 14, respectively and the relations $\|x^{k+1} - x^k\| = \|J_{\lambda T}u^{k+1} - J_{\lambda T}u^k\| \le \|u^{k+1} - u^k\|$ and $\|Fx^k + \xi^k\| = \|S_\lambda u^k\|$ for $\xi^k := \frac{1}{\lambda}(u^k - x^k) \in Tx^k$.

Next, we have $\lim_{k\to\infty}\|u^k - u^\star\|$ exists almost surely for all $u^\star \in \mathrm{zer}(S_\lambda)$ as in Theorem 14. Moreover, we also have $\lim_{k\to\infty}\|S_\lambda u^k\| = 0$ almost surely as in Theorem 14. Applying Lemma 31 again, we conclude that $\{u^k\}$ almost surely converges to a $\mathrm{zer}(S_\lambda)$-valued random variable $\bar{u}^\star \in \mathrm{zer}(S_\lambda)$.

Finally, since $x^k = J_{\lambda T}u^k$, $x^\star = J_{\lambda T}u^\star \in \mathrm{zer}(\Phi)$ almost surely, and $J_{\lambda T}$ is continuous (since it is nonexpansive), it is easy to show that $\{x^k\}$ also converges to $x^\star$ almost surely by the classical Continuous Mapping Theorem (Durrett, 2019, Exercise 1.3.3). ■

References

S. Adly and H. Attouch. First-order inertial algorithms involving dry friction damping. *Math. Program.*, pages 1–41, 2021.

R. P. Agarwal, M. Meehan, and D. O'regan. *Fixed point theory and applications*, volume 141. Cambridge university press, 2001.

A. Alacaoglu and Y. Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

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

J. K. Alcala, Y. T. Chow, and M. Sunkula. Moving anchor extragradient methods for smooth structured minimax problems. *arXiv preprint arXiv:2308.12359*, 2023.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

H. Attouch and A. Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Math. Program.*, 184(1):243–287, 2020.

H. Attouch and J. Fadili. From the Ravine method to the Nesterov method and vice versa: A dynamical system perspective. *SIAM J. Optim.*, 32(3):2074–2101, 2022.

H. Attouch and J. Peypouquet. Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. *Math. Program.*, 174(1-2):391–432, 2019.

H. Attouch, J. Peypouquet, and P. Redont. Backward–forward algorithms for structured monotone inclusions in Hilbert spaces. *J. Math. Anal. Appl.*, 457(2):1095–1117, 2018.

H. H. Bauschke and P. Combettes. *Convex analysis and monotone operators theory in Hilbert spaces*. Springer-Verlag, 2nd edition, 2017.

H. H. Bauschke, W. M. Moursi, and X. Wang. Generalized monotone operators and their averaged resolvents. *Math. Program.*, pages 1–20, 2020.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM J. Optim.*, 12:79–108, 2001.

A. Beznosikov, E. Gorbunov, H. Berard, and N. Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 172–235. PMLR, 2023.

A. Bohm, M. Sedlmayer, R. E. Csetnek, and R. I. Bot. Two steps at a time—Taking GAN training in stride with Tseng's method. *SIAM Journal on Mathematics of Data Science*, 4(2):750–771, 2022.

R. I. Bot and D. K. Nguyen. Fast Krasnoselśkii-Mann algorithm with a convergence rate of the fixed point iteration of $o(1/k)$. *SIAM Journal on Numerical Analysis*, 61(6):2813–2843, 2023.

R. I. Bot, P. Mertikopoulos, M. Staudigl, and P. T. Vuong. On the convergence of stochastic forward-backward-forward algorithms with variance reduction in pseudo-monotone variational inequalities. *NIPS 2018-Workshop on Smooth Games, Optimization and Machine Learning*, 2018.

R. I. Boţ, E. Chenchene, and J. M. Fadili. Generalized fast krasnoselskii-mann method with preconditioners. *arXiv preprint arXiv:2411.18574*, 2024.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.*, 60(2):223–311, 2018.

R. S. Burachik and A. Iusem. *Set-Valued Mappings and Enlargements of Monotone Operators*. New York: Springer, 2008.

X. Cai, C. Song, C. Guzmán, and J. Diakonikolas. A stochastic Halpern iteration with variance reduction for stochastic monotone inclusion problems. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2022)*, 2022a.

X. Cai, A. Alacaoglu, and J. Diakonikolas. Variance reduced halpern iteration for finite-sum monotone inclusions. In *The 12th International Conference on Learning Representations (ICLR)*, pages 1–33, 2024.

Y. Cai and W. Zheng. Accelerated Single-Call Methods for Constrained Min-Max Optimization. In *The 11th International Conference on Learning Representations, ICLR 2023*. The Eleventh International Conference on Learning Representations, ICLR 2023, 2023.

Y. Cai, A. Oikonomou, and W. Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022b.

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

C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32:393–403, 2019.

Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Math. Program.*, 165(1):113–149, 2017.

P. Combettes and J.-C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing, pages 185–212. Springer-Velarg, 2011.

L. Condat and P. Richtárik. Murana: A generic framework for stochastic variance-reduced optimization. In *Mathematical and Scientific Machine Learning*, pages 81–96. PMLR, 2022.

S. Cui and U. Shanbhag. On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *Set-Valued and Variational Analysis*, 29(2):453–499, 2021.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, pages 15210–15219, 2019.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with Optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

D. Davis. SMART: The stochastic monotone aggregated root-finding algorithm. *arXiv preprint arXiv:1601.00698*, 2016.

D. Davis. Variance reduction for root-finding problems. *Math. Program.*, pages 1–36, 2022.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1646–1654, 2014.

Y. Demidovich, G. Malinovsky, I. Sokolov, and P. Richtárik. A guide through the zoo of biased SGD. *Advances in Neural Information Processing Systems*, 36:23158–23171, 2023.

J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pages 1428–1451. PMLR, 2020.

D. Driggs, M. J. Ehrhardt, and C.-B. Schönlieb. Accelerating variance-reduced stochastic gradient methods. *Math. Program.*, (online first):1–45, 2020.

D. Driggs, J. Liang, and C.-B. Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022.

W. Du, D. Xu, X. Wu, and H. Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

K. Emmanouilidis, R. Vidal, and N. Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In *International Conference on Artificial Intelligence and Statistics*, pages 3682–3690. PMLR, 2024.

B. Evens, P. Pas, P. Latafat, and P. Patrinos. Convergence of the preconditioned proximal point method and Douglas-Rachford splitting in the absence of monotonicity. *arXiv preprint arXiv:2305.03605*, 2023.

F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*, volume 1-2. Springer-Verlag, 2003.

F. Faghri, C. N. Vasconcelos, D. J. Fleet, F. Pedregosa, and N. L. Roux. Bridging the gap between adversarial robustness and optimization bias. *ICLR*, 2025.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer-Verlag, New York, 2001.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

E. Gorbunov, F. Hanzely, and P. Richt. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022a.

E. Gorbunov, A. Taylor, S. Horváth, and G. Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: The case of negative comonotonicity. *International Conference on Machine Learning*, pages 11614–11641. PMLR, 2023b.

R. M. Gower, P. Richtárik, and F. Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Math. Program.*, 188(1):135–192, 2021.

B. Grimmer, H. Lu, P. Worah, and V. Mirrokni. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Math. Program.*, 201(1-2): 373–407, 2023.

B. Halpern. Fixed points of nonexpanding maps. *Bull. Am. Math. Soc.*, 73(6):957–961, 1967.

F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pages 2082–2093, 2018.

Y. He and R.-D. Monteiro. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM J. Optim.*, 26(1):29–56, 2016.

E. Ho, A. Rajagopalan, A. Skvortsov, S. Arulampalam, and M. Piraveenan. Game theory in defence applications: A review. *Sensors*, 22(3):1032, 2022.

K. Huang, N. Wang, and S. Zhang. An accelerated variance reduced extra-point approach to finite-sum VI and optimization. *arXiv preprint arXiv:2211.03269*, 2022.

A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.*, 27(2):686–724, 2017.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

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

A. Kannan and U. V. Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Comput. Optim. Appl.*, 74(3):779–820, 2019.

M. Khalafi and D. Boob. Accelerated primal-dual methods for convex-strongly-concave saddle point problems. In *International Conference on Machine Learning*, pages 16250–16270. PMLR, 2023.

D. Kim. Accelerated proximal point method for maximally monotone operators. *Math. Program.*, 190:57–87, 2021.

O. Kolossoski and R. D. Monteiro. An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems. *Optim. Meth. Soft.*, 32(6):1244–1272, 2017.

I. Konnov. *Combined relaxation methods for variational inequalities.* Springer-Verlag, 2001.

G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation. *SIAM J. Optim.*, 32(3):2041–2073, 2022.

D. Kovalev, S. Horvath, and P. Richtarik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020.

D. Kuhn, S. Shafiee, and W. Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

H. W. Kuhn, J. Harsanyi, R. Selten, J. Weibul, and E. van Damme. The work of John nash in game theory. *journal of economic theory*, 69(1):153–185, 1996.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.

S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPs2021)*, 2021a.

S. Lee and D. Kim. Semi-anchored multi-step gradient descent ascent method for structured nonconvex-nonconcave composite minimax problems. *arXiv preprint arXiv:2105.15042*, 2021b.

B. Li, M. Ma, and G. B. Giannakis. On the convergence of SARAH and beyond. *International Conference on Artificial Intelligence and Statistics*, pages 223–233, 2020, PMLR.

Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. *International Conference on Machine Learning*, pages 6286–6295, 2021, PMLR.

F. Lieder. On the convergence rate of the halpern-iteration. *Optim. Letters*, 15(2):405–418, 2021.

N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

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

P.-E. Maingé. Accelerated proximal algorithms with a correction term for monotone inclusions. *Applied Mathematics & Optimization*, 84(2):2027–2061, 2021.

P. E. Maingé. Fast convergence of generalized forward-backward algorithms for structured monotone inclusions. *J. Convex Anal.*, 29:893–920, 2022.

N. Martinez, M. Bertran, and G. Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

H. Namkoong and J. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

A. Nemirovski. Mini-course on convex programming algorithms. *Lecture notes*, 2013.

Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983. Translated as Soviet Math. Dokl.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2004.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2613–2621, 2017.

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

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

Z. Peng, Y. Xu, M. Yan, and W. Yin. ARock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM J. Scientific Comput.*, 38(5):2851–2879, 2016.

T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher. Solving stochastic weak Minty variational inequalities without increasing batch size. In *Proceedings of International Conference on Learning Representations (ICLR)*, pages 1–34, 2023.

H. N. Pham, M. L. Nguyen, T. D. Phan, and Q. Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res.*, 21:1–48, 2020.

R. R. Phelps. *Convex functions, monotone operators and differentiability*, volume 1364. Springer, 2009.

H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

R. Rockafellar and R. Wets. *Variational Analysis*, volume 317. Springer, 2004.

E. Ryu and W. Yin. *Large-scale convex optimization: Algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

E. K. Ryu and S. Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15 (1):3–43, 2016.

S. Sabach and S. Shtern. A first order method for solving convex bilevel optimization problems. *SIAM J. Optim.*, 27(2):640–660, 2017.

A. Sadiev, L. Condat, and P. Richtárik. Stochastic proximal point methods for monotone inclusions under expected similarity. *arXiv preprint arXiv:2405.14255*, 2024.

M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

C. Shi, M. Uehara, J. Huang, and N. Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In *International Conference on Machine Learning*, pages 20057–20094. PMLR, 2022.

S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT Press, 2012.

G. Swamy, S. Choudhury, J. A. Bagnell, and S. Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pages 10022–10032. PMLR, 2021.

Q. Tran-Dinh. Extragradient-Type Methods with $\mathcal{O}(1/k)$-Convergence Rates for Co-Hypomonotone Inclusions. *J. Global Optim.*, pages 1–25, 2023.

Q. Tran-Dinh. From Halpern's fixed-point iterations to Nesterov's accelerated interpretations for root-finding problems. *Comput. Optim. Appl.*, 87(1):181–218, 2024a.

Q. Tran-Dinh. Variance-Reduced Fast Krasnoselkii-Mann Methods for Finite-Sum Root-Finding Problems. *arXiv preprint arXiv:2406.02413*, 2024b.

Q. Tran-Dinh. Variance-Reduced Forward-Reflected-Backward Splitting Methods for Nonmonotone Generalized Equations. *Forty-Second International Conference on Machine Learning (ICML)*, 2025.

Q. Tran-Dinh and Y. Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021.

Q. Tran-Dinh and Y. Luo. Randomized Block-Coordinate Optimistic Gradient Algorithms for Root-Finding Problems. *Math. Oper. Res.*, in press, 2025.

Q. Tran-Dinh, H. N. Pham, T. D. Phan, and M. L. Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *Preprint: arXiv:1905.05920*, 2019.

Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. *Math. Program.*, 191:1005–1071, 2022.

S. J. Wright. Optimization Algorithms for Data Analysis. *IAS/Park City Mathematics Series*, pages 1–49, 2017.

J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

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

F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic mirror-prox algorithms for stochastic cartesian variational inequalities: Randomized block coordinate and optimal averaging schemes. *Set-Valued and Variational Analysis*, 26:789–819, 2018.

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

Y.-X. Yuan and Y. Zhang. Symplectic Extra-gradient type method for solving general non-monotone inclusion problem. *arXiv preprint arXiv:2406.10793*, 2024.