

Variance-Reduced Forward-Reflected-Backward Splitting Methods for Nonmonotone Generalized Equations

Quoc Tran-Dinh¹

Abstract

We develop two novel stochastic variance-reduction methods to approximate solutions of a class of nonmonotone [generalized] equations. Our algorithms leverage a new combination of ideas from the forward-reflected-backward splitting method and a class of unbiased variance-reduced estimators. We construct two new stochastic estimators within this class, inspired by the well-known SVRG and SAGA estimators. These estimators significantly differ from existing approaches used in minimax and variational inequality problems. By appropriately choosing parameters, both algorithms achieve state-of-the-art oracle complexity of $\mathcal{O}(n + n^{2/3}\epsilon^{-2})$ for obtaining an ϵ -solution in terms of the operator residual norm for a class of nonmonotone problems, where n is the number of summands and ϵ signifies the desired accuracy. This complexity aligns with the best-known results in SVRG and SAGA methods for stochastic nonconvex optimization. We test our algorithms on some numerical examples and compare them with existing methods. The results demonstrate promising improvements offered by the new methods compared to their competitors.

1. Introduction

[Non]linear equations and inclusions are cornerstones of computational mathematics, finding applications in diverse fields like engineering, mechanics, economics, statistics, optimization, and machine learning, see, e.g., (Bauschke & Combettes, 2017; Burachik & Iusem, 2008; Facchinei & Pang, 2003; Phelps, 2009; Ryu & Yin, 2022; Ryu & Boyd, 2016). These problems, known as *generalized equations* (Rockafellar & Wets, 1997), are equivalent to *fixed-point problems*. The recent revolution in modern machine learn-

ing and robust optimization has brought renewed interest to generalized equations and their special case: minimax problem. They serve as powerful tools for handling Nash's equilibria and minimax models in generative adversarial nets, adversarial training, and robust learning, see (Arjovsky et al., 2017; Goodfellow et al., 2014; Madry et al., 2018; Namkoong & Duchi, 2016). Notably, most problems arising from these applications are nonmonotone, nonsmooth, and large-scale. This paper develops new and simple stochastic algorithms with variance reduction for solving this class of problems, equipped with rigorous theoretical guarantees.

1.1. Nonmonotone finite-sum generalized equations

The central problem we study in this paper is the following [possibly nonmonotone] *generalized equation* (also known as *composite inclusion*) (Rockafellar & Wets, 1997):

$$\text{Find } x^* \in \mathbb{R}^p \text{ such that: } 0 \in Gx^* + Tx^*, \quad (\text{NI})$$

where $G : \mathbb{R}^p \rightarrow \mathbb{R}^p$ is a single-valued operator, possibly nonlinear, and $T : \mathbb{R}^p \rightrightarrows 2^{\mathbb{R}^p}$ is a multivalued mapping from \mathbb{R}^p to $2^{\mathbb{R}^p}$ (the set of all subsets of \mathbb{R}^p). In addition, we assume that G is given in the following *large finite-sum*:

$$Gx := \frac{1}{n} \sum_{i=1}^n G_i x, \quad (1)$$

where $G_i : \mathbb{R}^p \rightarrow \mathbb{R}^p$ are given operators for all $i \in [n] := \{1, 2, \dots, n\}$ and $n \gg 1$. This structure often arises from statistical learning, [generative] machine learning, networks, distributed systems, and data science. For simplicity of notation, we denote $\Psi := G + T$ and $\text{dom}(\Psi) := \text{dom}(G) \cap \text{dom}(T)$, where $\text{dom}(R)$ is the domain of R .

We highlight that the methods developed in this paper can be straightforwardly extended to tackle $Gx = \mathbb{E}_{\xi \sim \mathbb{P}} [\mathbf{G}(x, \xi)]$ as the expectation of a stochastic operator \mathbf{G} involving a random vector ξ defined on a probability space $(\Omega, \mathbb{P}, \Sigma)$.

1.2. Equivalent forms and special cases

The model (NI) covers many fundamental problems in optimization and related fields, including the following ones.

(a) **[Non]linear equation.** If $T = 0$, then (NI) reduces to the following [non]linear equation:

$$\text{Find } x^* \in \text{dom}(G) \text{ such that: } Gx^* = 0. \quad (\text{NE})$$

¹Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. Correspondence to: Quoc Tran-Dinh <quocdt@email.unc.edu>.

Both (NI) and (NE) are also called *root-finding problems*. Clearly, (NE) is a special case of (NI). However, under appropriate assumptions on G and/or T (e.g., using the resolvent of T), one can also transform (NI) to (NE). Let $\text{zer}(\Psi) := \{x^* \in \text{dom}(\Psi) : 0 \in \Psi x^*\}$ and $\text{zer}(G) := \{x^* \in \text{dom}(G) : Gx^* = 0\}$ be the solution sets of (NI) and (NE), respectively, which are assumed to be nonempty.

(b) **Variational inequality problem (VIP).** If $T = \mathcal{N}_{\mathcal{X}}$, the normal cone of a nonempty, closed, and convex set \mathcal{X} in \mathbb{R}^p , then (NI) reduces to the following VIP:

Find $x^* \in \mathcal{X}$ such that: $\langle Gx^*, x - x^* \rangle \geq 0, \forall x \in \mathcal{X}$. (VIP)

If $T = \partial g$, the subdifferential of a convex function g , then (NI) reduces to a mixed VIP, denoted by MVIP. Both VIP and MVIP cover many problems in practice, including minmax problems and Nash's equilibria, see, e.g., (Burachik & Iusem, 2008; Facchinei & Pang, 2003; Phelps, 2009).

(c) **Minimax optimization.** Another important special case of (NI) (or MVIP) is the following minimax optimization (or saddle-point) problem, which has found various applications in machine learning and robust optimization:

$$\min_{u \in \mathbb{R}^{p_1}} \max_{v \in \mathbb{R}^{p_2}} \left\{ \mathcal{L}(u, v) := \varphi(u) + \mathcal{H}(u, v) - \psi(v) \right\}, \quad (\text{SP})$$

where $\mathcal{H} : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \rightarrow \mathbb{R}$ is a smooth function, and φ and ψ are proper, closed, and convex. Let us define $x := [u, v] \in \mathbb{R}^p$ as the concatenation of u and v with $p := p_1 + p_2$, $Gx := [\nabla_u \mathcal{H}(u, v), -\nabla_v \mathcal{H}(u, v)]$, and $Tx := [\partial \varphi(u), \partial \psi(v)]$. Then, the optimality condition of (SP) is written in the form of (NI). Since (VIP), and in particular, (SP) are special cases of (NI), our algorithms for (NI) in the sequel can be specified to solve these problems.

(d) **Fixed-point problem.** Problem (NE) is equivalent to the following fixed-point problem:

$$\text{Find } x^* \in \text{dom}(F) \text{ such that: } x^* = Fx^*, \quad (\text{FP})$$

where $F := \mathbb{I} - \lambda G$ with \mathbb{I} being the identity operator and $\lambda > 0$. Since (FP) is equivalent to (NE), our algorithms for (NE) from this paper can also be applied to solve (FP).

1.3. Motivation

Our work is mainly motivated by the following aspects.

(i) *Recent applications.* Both (NE) and (NI) cover the minmax problem (SP) as a special case. This minimax problem, especially in nonconvex-nonconcave settings, has recently gained its popularity as it provides a powerful tool to model applications in generative adversarial networks (Arjovsky et al., 2017; Goodfellow et al., 2014), robust and distributionally robust optimization (Ben-Tal et al., 2009; Bertsimas & Caramanis, 2011; Levy et al., 2020), adversarial training (Madry et al., 2018), online optimization (Bhatia & Sridharan, 2020), and reinforcement learning (Azar et al., 2017; Zhang et al., 2021). Our work is motivated by those applications.

(ii) *Optimality certification.* Existing stochastic methods often target special cases of (NI) such as (NE) and (VIP). In addition, these methods frequently rely on a monotonicity assumption, which excludes many problems of current interest, e.g., (Alacaoglu et al., 2023; Alacaoglu & Malitsky, 2022; Beznosikov et al., 2023; Gorbunov et al., 2022a; Loizou et al., 2021). Furthermore, existing methods analyze convergence based on a [duality] **gap function** (Facchinei & Pang, 2003) or a **restricted gap function** (Nesterov, 2007). As discussed in (Cai et al., 2024; Diakonikolas, 2020), these metrics have limitations, particularly in nonmonotone settings. It is important to note that standard gap functions are not applicable to our settings due to Assumption 1.4. Regarding oracle complexity, several works, e.g., (Alacaoglu & Malitsky, 2022; Beznosikov et al., 2023; Gorbunov et al., 2022a; Loizou et al., 2021) claim an oracle complexity of $\mathcal{O}(n + \sqrt{n}\epsilon^{-1})$ to attain an ϵ -solution, but this is measured using a restricted gap function. Again, as highlighted in (Cai et al., 2024; Diakonikolas, 2020), this certification does not translate to the operator residual norm and is inapplicable to nonmonotone settings. Therefore, a direct comparison between our results and these previous works is challenging due to these methodological discrepancies (see Table 1).

(iii) *New and simple algorithms.* Various existing stochastic methods for solving (VIP) and (NI) rely on established techniques. These include mirror-prox/averaging and extragradient-type schemes combined with the classic Robbin-Monro stochastic approximation (Robbins & Monro, 1951) (e.g., (Cui & Shanbhag, 2021; Iusem et al., 2017; Juditsky et al., 2011; Kannan & Shanbhag, 2019; Kotsalis et al., 2022; Yousefian et al., 2018)). Some approaches utilize increasing mini-batch sizes for variance reduction (e.g., (Iusem et al., 2017; Pethick et al., 2023)). Recent works have explored alternative variance-reduced methods for (NI) and its special cases (e.g., (Alacaoglu et al., 2023; Alacaoglu & Malitsky, 2022; Bot et al., 2019; Cai et al., 2022; Davis, 2022)). However, these methods primarily adapt existing optimization estimators to approximate the operator G without significant differences. Our approach departs from directly approximating G . Instead, we construct an intermediate quantity $S_\gamma^k := Gx^k - \gamma Gx^{k-1}$ as a linear combination of two consecutive evaluations of G (i.e. Gx^k and Gx^{k-1}). We then develop stochastic variance-reduced estimators specifically for S_γ^k . Note that S_γ^k alone is not new, but our idea of using it in stochastic methods is new. This idea allows us to design new and simple algorithms with a single loop for both (NE) and (NI) where the state-of-the-art oracle complexity is achieved (cf. Sections 3 and 4).

1.4. Basic assumptions

We tackle both (NE) and (NI) covered by the following basic assumptions (see (Bauschke & Combettes, 2017) for terminologies and concepts used in these assumptions).

Assumption 1.1. [Well-definedness] $\text{zer}(\Psi)$ of (NI) and

$\text{zer}(G)$ of (NE) are nonempty.

Assumption 1.2. [Maximal monotonicity of T] T in (NI) is maximally monotone on $\text{dom}(T)$.

Assumption 1.3. [Lipschitz continuity of G] G in (1) is L -averaged Lipschitz continuous, i.e. $\forall x, y \in \text{dom}(G)$:

$$\frac{1}{n} \sum_{i=1}^n \|G_i x - G_i y\|^2 \leq L^2 \|x - y\|^2. \quad (2)$$

Assumption 1.4. [Weak-Minty solution] There exist a solution $x^* \in \text{zer}(\Psi)$ and $\kappa \geq 0$ such that $\langle Gx + v, x - x^* \rangle \geq -\kappa \|Gx + v\|^2$ for all $x \in \text{dom}(\Psi)$ and $v \in Tx$.

While Assumption 1.1 is basic, Assumption 1.2 guarantees the single-valued and well-definiteness of the resolvent J_T of T . In fact, this assumption can be relaxed to some classes of nonmonotone operators T , but we omit this extension. The L -averaged Lipschitz continuity (2) is standard and has been used in most deterministic, randomized, and stochastic methods. It is slightly stronger than the L -Lipschitz continuity of the sum G . The star-co-hypomonotonicity in Assumption 1.4 is significantly different from the star-strong monotonicity used in, e.g., (Kotsalis et al., 2022). It covers a class of nonmonotone operators G (see Supp. Doc. A.2 for a concrete example). It is also weaker than the co-hypomonotonicity, used, e.g., in (Cai et al., 2024).

1.5. Contribution and related work

Our primary goal is to develop a class of variance-reduction methods to solve both (NE) and (NI), their special cases such as (VIP) and (SP), and equivalent problems, like (FP).

Our contribution. Our main contribution consists of:

- (a) We exploit the variable S_γ^k in (FRQ) and introduce a class of unbiased variance-reduced estimators \tilde{S}_γ^k for S_γ^k , not for G , covered by Definition 2.1.
- (b) We construct two instances of \tilde{S}_γ^k by leveraging the SVRG (Johnson & Zhang, 2013) and SAGA (Defazio et al., 2014) estimators, respectively that fulfill our Definition 2.1. These estimators are also of independent interest, and can be applied to develop other methods.
- (c) We develop a variance-reduced forward-reflected-type method (VFR) to solve (NE) required $\mathcal{O}(n + n^{2/3}\epsilon^{-2})$ evaluations of G_i to obtain an ϵ -solution.
- (d) We design a novel stochastic variance-reduced forward-reflected-backward splitting method (VFRBS) to solve (NI), also required $\mathcal{O}(n + n^{2/3}\epsilon^{-2})$ evaluations of G_i .

Table 1 below compares our work and some existing single-loop variance-reduction methods, but in either the co-coercive or monotone settings. Now, let us highlight the following points of our contribution. First, our intermediate quantity S_γ^k can be viewed as a generalization of the forward-reflected-backward splitting (FRBS) operator (Malitsky & Tam, 2020) or an optimistic gradient operator (Daskalakis et al., 2018) used in the literature. However, the chosen

range $\gamma \in (1/2, 1)$ excludes these classical methods from recovering as special cases of S_γ^k . Second, since our SVRG and SAGA estimators are designed specifically for S_γ^k , they differ from existing estimators in the literature, including recent works (Alacaoglu et al., 2023; Alacaoglu & Malitsky, 2022; Bot et al., 2019). Third, both proposed algorithms are single-loop and straightforward to implement. Fourth, our algorithm for nonlinear inclusions (NI) significantly differs from existing methods, including deterministic ones, due to the additional term $\gamma^{-1}(2\gamma - 1)(y^k - x^k)$. For a comprehensive survey of deterministic methods, we refer to (Tran-Dinh, 2023). Fifth, our oracle complexity estimates rely on the metric $\mathbb{E}[\|Gx^k\|^2]$ or $\mathbb{E}[\|Gx^k + v^k\|^2]$ for $v^k \in Tx^k$, commonly used in nonmonotone settings. Unlike the monotone case, this metric cannot be directly converted to a gap function, see, e.g., (Alacaoglu et al., 2023; Alacaoglu & Malitsky, 2022). Our complexity bounds match the best known in stochastic nonconvex optimization using SAGA or SVRG without additional enhancements, e.g., utilizing a nested technique as in (Zhou et al., 2018).

Related work. Since both theory and solution methods for solving (NE) and (NI) are ubiquitous, see, e.g., (Bauschke & Combettes, 2017; Burachik & Iusem, 2008; Facchinei & Pang, 2003; Phelps, 2009; Ryu & Yin, 2022; Ryu & Boyd, 2016), especially under the monotonicity, we only highlight the most recent related works (see more in Supp. Doc. A).

(i) *Weak-Minty solution.* Assumption 1.4 is known as a weak-Minty solution of (NI) (in particular, of (NE)), which has been widely used in recent works, e.g., (Böhm, 2022; Diakonikolas et al., 2021; Lee & Kim, 2021; Pethick et al., 2022; Tran-Dinh, 2023a) for deterministic methods and, e.g., (Lee & Kim, 2021; Pethick et al., 2023; Tran-Dinh & Luo, 2025) for stochastic methods. This weak-Minty solution condition is weaker than the co-hypomonotonicity (Bauschke et al., 2020), which was used earlier in proximal-point methods (Combettes & Pennanen, 2004). Diakonikolas et al. exploited this condition to develop an extragradient variant (called EG+) to solve (NE). Following up works include (Böhm, 2022; Cai & Zheng, 2023; Luo & Tran-Dinh, 2022; Pethick et al., 2022; Tran-Dinh, 2023a). A recent survey in (Tran-Dinh, 2023) provides several deterministic methods that rely on this condition. This assumption covers a class of nonmonotone operators G or $G + T$.

(ii) *Stochastic approximation methods.* Stochastic methods for both (NE) and (NI) and their special cases have been extensively developed, see, e.g., (Juditsky et al., 2011; Kotsalis et al., 2022; Pethick et al., 2023). Several methods exploited mirror-prox and averaging techniques such as (Juditsky et al., 2011; Kotsalis et al., 2022), while others relied on projection or extragradient schemes, e.g., (Cui & Shanbhag, 2021; Iusem et al., 2017; Kannan & Shanbhag, 2019; Pethick et al., 2023; Yousefian et al., 2018). Many of these algorithms use standard Robbin-Monro stochastic

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

Papers	Problem	Assumptions	Estimators	Residual Rates	Oracle Complexity
(Davis, 2022)	(NE) and (NI)	co-coercive/SQM	SVRG & SAGA	linear	$\mathcal{O}((L/\mu) \log(\epsilon^{-1}))$
(Tran-Dinh, 2024)	(NE) and (NI)	co-coercive	a class	$\mathcal{O}(1/k^2)$	$\mathcal{O}(n + n^{2/3}\epsilon^{-1})$
(Cai et al., 2024)	(NE) and (NI)	co-coercive	SARAH	$\mathcal{O}(1/k^2)$	$\mathcal{O}(n + n^{1/2} \log(n)\epsilon^{-1})$
(Alacaoglu & Malitsky, 2022)	(VIP)	monotone	SVRG	\mathbf{x}	$\mathcal{O}(n + n^{1/2}\epsilon^{-1})$
(Alacaoglu et al., 2023)	(VIP)	monotone	SVRG	\mathbf{x}	$\mathcal{O}(n\epsilon^{-1})$
Ours	(NE) and (NI)	weak Minity	a class	$\mathcal{O}(1/k)$	$\mathcal{O}(n + n^{2/3}\epsilon^{-2})$

Notes: **SQM** means “strong quasi-monotonicity”; **Residual Rate** is the convergence rate on $\mathbb{E}[\|Gx^k\|^2]$ or $\mathbb{E}[\|Gx^k + v^k\|^2]$ for $v^k \in Tx^k$; and **a class** is a class of variance-reduced estimators. The complexity of (Alacaoglu & Malitsky, 2022) and (Alacaoglu et al., 2023) marked by **magenta** is on a gap function, a different metric than in other works in Table 1. Thus, it is unclear how to compare them.

approximation with fixed or increasing batch sizes. Some other works generalized the analysis to a general class of algorithms such as (Beznosikov et al., 2023; Gorbunov et al., 2022a; Loizou et al., 2021) covering both standard stochastic approximation and variance reduction algorithms.

(iii) *Variance-reduction methods.* Variance-reduction techniques have been broadly explored in optimization, where many estimators were proposed, including SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013), SARAH (Nguyen et al., 2017), and Hybrid-SGD (Tran-Dinh et al., 2019; 2022), and STORM (Cutkosky & Orabona, 2019). Researchers have adopted these estimators to develop methods for (NE) and (NI). For example, (Davis, 2022) proposed a SAGA-type methods for (NE) under a [quasi]-strong monotonicity. The authors in (Alacaoglu et al., 2023; Alacaoglu & Malitsky, 2022) employed SVRG estimators and developed methods for (VIP). Other works can be found in (Bot et al., 2019; Carmon et al., 2019; Chavdarova et al., 2019; Huang et al., 2022; Palaniappan & Bach, 2016; Yu et al., 2022). All of these results are different from ours. Some recent works exploited Halpern’s fixed-point iterations and develop corresponding variance-reduced methods, see, e.g., (Cai et al., 2024; 2022). However, varying parameters or incorporating double-loop/inexact methods must be used to achieve improved theoretical oracle complexity. We believe that such approaches may be challenging to select parameters and to implement in practice. Finally, unlike optimization, it has been realized that using biased estimators such as SARAH or Hybrid-SGD/STORM for (NI) (including (SP) and (VIP)) is challenging due to the lack of an objective function, a key metric to prove convergence, and product terms like $\langle e^k, x^{k+1} - x^* \rangle$ in convergence analyses, where e^k is a bias rendered from \tilde{S}_γ^k (see Supp. Doc. A).

Notation. We use $\mathcal{F}_k := \sigma(x^0, x^1, \dots, x^k)$ to denote the σ -algebra generated by x^0, \dots, x^k up to the iteration k . $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_k]$ denotes the conditional expectation w.r.t. \mathcal{F}_k , and $\mathbb{E}[\cdot]$ is the total expectation. We also use $\mathcal{O}(\cdot)$ to characterize convergence rates and oracle complexity. For an operator G , $\text{dom}(G) := \{x : Gx \neq \emptyset\}$ denotes its domain, and J_G denotes its resolvent.

Paper organization. Section 2 introduces S_γ^k and defines a

class of stochastic estimators for it. It also constructs two instances: SVRG and SAGA, and proves their key properties. Section 3 develops an algorithm for solving (NE) and establishes its oracle complexity. Section 4 designs a new algorithm for solving (NI) and proves its oracle complexity. Section 5 presents two concrete numerical examples. Proofs and additional results are deferred to Supp. Docs. A to E.

2. Forward-Reflected Quantity and Its Stochastic Variance-Reduced Estimators

We first define our forward-reflected quantity (FRQ) for G in (NE) and (NI) using here. Next, we propose a class of unbiased variance-reduced estimators for FRQ. Finally, we construct two instances relying on the two well-known estimators: SVRG from (Johnson & Zhang, 2013) and SAGA from (Defazio et al., 2014).

2.1. The forward-reflected quantity

Our methods for solving (NE) and (NI) rely on the following intermediate quantity constructed from G via two consecutive iterates x^{k-1} and x^k controlled by $\gamma \in [0, 1]$:

$$S_\gamma^k := Gx^k - \gamma Gx^{k-1}. \quad (\text{FRQ})$$

Here, $\gamma \in (\frac{1}{2}, 1)$ plays a crucial role in our methods in the sequel. Clearly, if $\gamma = \frac{1}{2}$, then we can write $S_{1/2}^k = \frac{1}{2}Gx^k + \frac{1}{2}(Gx^k - Gx^{k-1}) = \frac{1}{2}[2Gx^k - Gx^{k-1}]$ used in both the forward-reflected-backward splitting (FRBS) method (Malitsky & Tam, 2020) and the optimistic gradient method (Daskalakis et al., 2018). In deterministic unconstrained settings (i.e. solving (NE)), see (Tran-Dinh, 2023), FRBS is also equivalent to Popov’s past-extragradient method (Popov, 1980), reflected-forward-backward splitting algorithm (Cevher & Vũ, 2021; Malitsky, 2015), and optimistic gradient scheme (Daskalakis et al., 2018). In the deterministic constrained case, i.e. solving (NI), these methods are different. Since $\gamma \in (\frac{1}{2}, 1)$, our methods below exclude these classical schemes. However, due to a similarity pattern of (FRQ) and FRBS, we still term our quantity S_γ^k by the “**forward-reflected quantity**”, abbreviated by FRQ.

2.2. Unbiased variance-reduced estimators for FRQ

Now, let us propose the following class of stochastic variance-reduced estimators \tilde{S}_γ^k of S_γ^k .

Definition 2.1. A stochastic estimator \tilde{S}_γ^k is said to be a *stochastic unbiased variance-reduced estimator* of S_γ^k in (FRQ) if there exist constants $\rho \in (0, 1]$, $C \geq 0$ and $\hat{C} \geq 0$, and a nonnegative sequence $\{\Delta_k\}$ such that:

$$\begin{cases} \mathbb{E}_k[\tilde{S}_\gamma^k - S_\gamma^k] = 0, \\ \mathbb{E}[\|\tilde{S}_\gamma^k - S_\gamma^k\|^2] \leq \Delta_k, \\ \Delta_k \leq (1 - \rho)\Delta_{k-1} + C \cdot U_k + \hat{C} \cdot U_{k-1}, \end{cases} \quad (3)$$

where $U_k := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - G_i x^{k-1}\|^2]$.

Here, $\Delta_{-1} = 0$, $x^{-2} = x^{-1} = x^0$, and $\mathbb{E}_k[\cdot]$ and $\mathbb{E}[\cdot]$ are the conditional and total expectations defined earlier, respectively. The condition $\rho > 0$ is important to achieve a variance reduction as long as x^k is close to x^{k-1} and x^{k-1} is close to x^{k-2} . Otherwise, \tilde{S}_γ^k may not be a variance-reduced estimator of S_γ^k . Since S_γ^k is evaluated at both x^{k-1} and x^k , our bound for the estimator \tilde{S}_γ^k depends on three consecutive points x^{k-2} , x^{k-1} , and x^k , which is different from previous works, including (Alacaoglu et al., 2021; Beznosikov et al., 2023; Davis, 2022; Driggs et al., 2020).

Now, we will construct two variance-reduced estimators satisfying Definition 2.1 by exploiting SVRG (Johnson & Zhang, 2013) and SAGA (Defazio et al., 2014).

(a) **Loopless-SVRG estimator for S_γ^k .** Consider a mini-batch $\mathcal{B}_k \subseteq [n] := \{1, 2, \dots, n\}$ with a fixed batch size $b := |\mathcal{B}_k|$. Denote $G_{\mathcal{B}_k} z := \frac{1}{b} \sum_{i \in \mathcal{B}_k} G_i z$ for $z \in \text{dom}(G)$. We define the following estimator for S_γ^k :

$$\begin{aligned} \tilde{S}_\gamma^k &:= (1 - \gamma)(Gw^k - G_{\mathcal{B}_k} w^k) \\ &\quad + G_{\mathcal{B}_k} x^k - \gamma G_{\mathcal{B}_k} x^{k-1}, \end{aligned} \quad (\text{L-SVRG})$$

where the snapshot point w^k is selected randomly as follows:

$$w^{k+1} := \begin{cases} x^k & \text{with probability } \mathbf{p} \\ w^k & \text{with probability } 1 - \mathbf{p}. \end{cases} \quad (4)$$

The probability $\mathbf{p} \in (0, 1)$ will appropriately be chosen later by nonuniformly flipping a coin. This estimator is known as a loopless variant (Kovalev et al., 2020) of the SVRG estimator (Johnson & Zhang, 2013). However, it is different from existing ones used in root-finding algorithms, including (Davis, 2022) because we define it for S_γ^k , not for Gx^k . In addition, the first term is also damped by a factor $1 - \gamma$ to guarantee the unbiasedness of \tilde{S}_γ^k to S_γ^k .

The following lemma shows that \tilde{S}_γ^k satisfies Definition 2.1.

Lemma 2.2. Let S_γ^k be given by (FRQ) and \tilde{S}_γ^k be generated by the SVRG estimator (L-SVRG) and

$$\Delta_k := \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1 - \gamma)G_i w^k\|^2].$$

Then, \tilde{S}_γ^k satisfies Definition 2.1 with this $\{\Delta_k\}$, $\rho := \frac{\mathbf{p}}{2}$, $C := \frac{4-6\mathbf{p}+3\mathbf{p}^2}{b\mathbf{p}}$, and $\hat{C} := \frac{2\gamma^2(2-3\mathbf{p}+\mathbf{p}^2)}{b\mathbf{p}}$.

(b) **SAGA estimator for S_γ^k .** Given S_γ^k as in (FRQ) and a mini-batch estimator $G_{\mathcal{B}_k}$ as in (L-SVRG), we construct the following SAGA estimator for S_γ^k :

$$\begin{aligned} \tilde{S}_\gamma^k &:= [G_{\mathcal{B}_k} x^k - \gamma G_{\mathcal{B}_k} x^{k-1} - (1 - \gamma)\hat{G}_{\mathcal{B}_k}^k] \\ &\quad + \frac{1-\gamma}{n} \sum_{i=1}^n \hat{G}_i^k, \end{aligned} \quad (\text{SAGA})$$

where \mathcal{B}_k is a mini-batch of size b , and \hat{G}_i^k is updated as

$$\hat{G}_i^{k+1} := \begin{cases} G_i x^k & \text{if } i \in \mathcal{B}_k, \\ \hat{G}_i^k & \text{otherwise.} \end{cases} \quad (5)$$

To form \tilde{S}_γ^k , we need to store n components \hat{G}_i^k computed so far for $i \in [n]$ in a table $\mathcal{T}_k := [\hat{G}_1^k, \hat{G}_2^k, \dots, \hat{G}_n^k]$ initialized at $\hat{G}_i^0 := G_i x^0$ for all $i \in [n]$. Clearly, the SAGA estimator requires significant memory to store \mathcal{T}_k if n and p are both large. We have the following result.

Lemma 2.3. Let S_γ^k be defined by (FRQ) and \tilde{S}_γ^k be generated by the SAGA estimator (SAGA), and

$$\Delta_k := \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1 - \gamma)\hat{G}_i^k\|^2].$$

Then, \tilde{S}_γ^k satisfies Definition 2.1 with this $\{\Delta_k\}$, $\rho := \frac{b}{2n} \in (0, 1]$, $C := \frac{[2(n-b)(2n+b)+b^2]}{nb}$, and $\hat{C} := \frac{2(n-b)(2n+b)\gamma^2}{nb}$.

We only provide two instances: (L-SVRG) and (SAGA) covered by Definition 2.1. However, we believe that similar estimators for S_γ^k relied on, e.g., JacSketch (Gower et al., 2021) or SEGA (Hanzely et al., 2018), among others can fulfill our Definition 2.1.

3. A Variance-Reduced Forward-Reflected Method for [Non]linear Equations

We first utilize the class of stochastic estimators in Definition 2.1 to develop a variance-reduced forward-reflected (VFR) method for solving (NE) under Ass. 1.3 and 1.4.

3.1. The VFR method and its convergence guarantee

(a) **VFR Method.** Our method is described as follows. Starting from $x^0 \in \text{dom}(G)$, at each iteration $k \geq 0$, we construct an estimator \tilde{S}_γ^k satisfying Definition 2.1 with parameters $\rho \in (0, 1]$, $C \geq 0$, and $\hat{C} \geq 0$, and then update

$$x^{k+1} := x^k - \eta \tilde{S}_\gamma^k, \quad (\text{VFR})$$

where $\eta > 0$ and $\gamma > 0$ are determined below, $x^{-1} = x^{-2} := x^0$, and $\tilde{S}_\gamma^0 := (1 - \gamma)Gx^0$.

At least two estimators \tilde{S}_γ^k : the Loopless-SVRG estimator in (L-SVRG) and the SAGA estimator in (SAGA), can be used in our method (VFR). In terms of *per-iteration complexity*, each iteration k of VFR, the loopless SVRG variant requires three mini-batch evaluations $G_{\mathcal{B}_k} w^k$, $G_{\mathcal{B}_k} x^k$, and $G_{\mathcal{B}_k} x^{k-1}$ of G , and occasionally computes one full evaluation Gw^k

of G with the probability \mathbf{p} . It needs one more mini-batch evaluation $G_{\mathcal{B}_k} x^{k-1}$ compared to SVRG-type methods in optimization. Similarly, the SAGA estimator also requires two mini-batch evaluations $G_{\mathcal{B}_k} x^k$ and $G_{\mathcal{B}_k} x^{k-1}$, which is one more mini-batch $G_{\mathcal{B}_k} x^{k-1}$ compared to SAGA-type methods in optimization, see, e.g., (Reddi et al., 2016b). The SAGA estimator can avoid the occasional full-batch evaluation Gw^k from L-SVRG, but as a compensation, we need to store a table $\mathcal{T}_k := [\hat{G}_1^k, \hat{G}_2^k, \dots, \hat{G}_n^k]$, which requires significant memory in the large-scale regime.

(b) **Convergence guarantee.** Fixed $\gamma \in (\frac{1}{2}, 1)$, with ρ, C , and \hat{C} as in Definition 2.1 we define

$$M := \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{1+6\gamma}{3(2\gamma-1)} \cdot \frac{C+\hat{C}}{\rho} \quad \text{and} \quad \delta := \frac{2\gamma-1}{8\sqrt{M}}. \quad (6)$$

Then, the following theorem states the convergence of (VFR), whose proof is given in Supp. Doc. C.

Theorem 3.1. *Let us fix $\gamma \in (\frac{1}{2}, 1)$, and define M and δ as in (6). Suppose that Assumptions 1.1, 1.3, and 1.4 hold for (NE) with some $\kappa \geq 0$ such that $L\kappa \leq \delta$. Let $\{x^k\}$ be generated by (VFR) using a learning rate $\eta > 0$ such that $\frac{8\kappa}{2\gamma-1} \leq \eta \leq \frac{1}{L\sqrt{M}}$. Then, the following bounds hold:*

$$\begin{aligned} \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] &\leq \frac{\Theta_1 \|x^0 - x^*\|^2}{K+1}, \\ \frac{1}{K+1} \sum_{k=1}^K \mathbb{E}[\|x^k - x^{k-1}\|^2] &\leq \frac{\Theta_2 \|x^0 - x^*\|^2}{K+1}, \end{aligned} \quad (7)$$

where $\Theta_1 := \frac{2(1+L^2\eta^2)}{\gamma(1-\gamma)\eta^2}$ and $\Theta_2 := \frac{8(1+L^2\eta^2)}{3(2\gamma-1)(1-ML^2\eta^2)}$

Theorem 3.1 only proves a $\mathcal{O}(1/K)$ convergence rate of both $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2]$ and $\frac{1}{K+1} \sum_{k=1}^K \mathbb{E}[\|x^k - x^{k-1}\|^2]$, but does not characterize the oracle complexity of (VFR). If we choose $\gamma := \frac{3}{4}$, then from (6), we have $M = \frac{57}{24} + \frac{11(C+\hat{C})}{3\rho}$ and $\delta = \frac{1}{16\sqrt{M}}$, which can simplify the bounds in Theorem 3.1. In addition, it allows $\kappa > 0$ such that $L\kappa \leq \delta = \mathcal{O}(\sqrt{\rho})$, which means that κ can be positive, but depends on $\sqrt{\rho}$. This condition allows us to cover a class of nonmonotone operators G , where a weak-Minty solution exists as stated in Assumption 1.4.

3.2. Complexity Bounds of VFR with SVRG and SAGA

Let us first apply Theorem 3.1 to the mini-batch SVRG estimator (L-SVRG) in Section 2. For simplicity, we choose $\gamma := \frac{3}{4}$ and $\eta := \frac{1}{L\sqrt{M}}$, but any $\gamma \in (\frac{1}{2}, 1)$ still works.

Corollary 3.2. *Suppose that Assumptions 1.1, 1.3, and 1.4 hold for (NE) with $\kappa \geq 0$ as in Theorem 3.1. Let $\{x^k\}$ be generated by (VFR) using (L-SVRG), $\gamma := \frac{3}{4}$, and $\eta := \frac{1}{L\sqrt{M}} \geq \frac{0.1440\sqrt{b}\mathbf{p}}{L}$, provided that $b\mathbf{p}^2 \leq 1$. Then*

$$\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \frac{526L^2\|x^0 - x^*\|^2}{b\mathbf{p}^2(K+1)}. \quad (8)$$

For $\epsilon > 0$, if we choose $\mathbf{p} := n^{-1/3}$ and $b := \lfloor n^{2/3} \rfloor$, then (VFR) requires $\mathcal{T}_{G_i} := n + \lfloor \frac{4\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \rfloor$ evaluations of G_i to attain $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \epsilon^2$, where $\Gamma := 731$.

Corollary 3.2 states that the oracle complexity of (VFR) is $\mathcal{O}(n + n^{2/3}\epsilon^{-2})$, matching (up to a constant) the one of SVRG in nonconvex optimization, see, e.g., (Allen-Zhu & Hazan, 2016; Reddi et al., 2016a). It improves by a factor $\mathcal{O}(n^{1/3})$ compared to deterministic counterparts. This complexity is known to be the best for SVRG so far without any additional enhancement (e.g., nested techniques (Zhou et al., 2018)) even for a special case of (NE): $Gx = \nabla f(x)$ in nonconvex optimization.

Note that η can be computed explicitly when b and \mathbf{p} are given. For example, if $n = 10000$ and we choose $\mathbf{p} = n^{-1/3} = 0.0464$ and $b = \lfloor n^{2/3} \rfloor = 464$, then $\eta = \frac{0.1456}{L}$. If $\mathbf{p} = 0.1$, then $\eta = \frac{0.3038}{L}$. Note that, in general, we can choose appropriate $p := \mathcal{O}(n^{-1/3})$ and $b := \mathcal{O}(n^{2/3})$.

Alternatively, we can apply Theorem 3.1 to (SAGA).

Corollary 3.3. *Suppose that Assumptions 1.1, 1.3, and 1.4 hold for (NE) with $\kappa \geq 0$ as in Theorem 3.1. Let $\{x^k\}$ be generated by (VFR) using (SAGA), $\gamma := \frac{3}{4}$, and $\eta := \frac{1}{L\sqrt{M}} \geq \frac{0.1494b^{3/2}}{nL}$, provided that $1 \leq b \leq n^{2/3}$. Then*

$$\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \frac{489L^2\|x^0 - x^*\|^2}{b\mathbf{p}^2(K+1)}. \quad (9)$$

Moreover, for a given $\epsilon > 0$, if we choose $b := \lfloor n^{2/3} \rfloor$, then (VFR) requires $\mathcal{T}_{G_i} := n + \lfloor \frac{3\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \rfloor$ evaluations of G_i to achieve $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \epsilon^2$, where $\Gamma := 2816$.

Similar to Corollary 3.2, the learning rate η in Corollary 3.3 can explicitly be computed if we know n and b . For instance, if $n = 10000$, and we choose $b = \lfloor n^{2/3} \rfloor$, then $\eta = \frac{0.1603}{L}$.

If $\kappa = 0$, i.e. G reduces to a star-monotone operator, then we can choose $\gamma \in (\frac{1}{2}, 1)$ and η as:

- For SVRG, we have $\eta \in (0, \frac{1}{L\sqrt{M}}]$. If we choose $\mathbf{p} = \mathcal{O}(n^{-1/3})$ and $b = \mathcal{O}(n^{2/3})$, then $\eta = \mathcal{O}(\frac{1}{L})$;
- For SAGA, we have $\eta \in (0, \frac{1}{L\sqrt{M}}]$. If we choose $b = \mathcal{O}(n^{2/3})$, then $\eta = \mathcal{O}(\frac{1}{L})$.

Hitherto, the constant factor Γ in both corollaries is still relatively large, but it can be further improved by refining our technical proofs (e.g., carefully using Young's inequality).

4. A New Variance-Reduced FRBS Method for Nonmonotone Generalized Equations

In this section, we develop a new stochastic variance-reduced forward-reflected-backward splitting (FRBS) method to solve (NI) under Assumptions 1.2, 1.3, and 1.4.

4.1. The algorithm and its convergence

(a) **The variance-reduced FRBS method (VFRBS).** Our scheme for solving (NI) is as follows. Starting from $x^0 \in \text{dom}(\Psi)$, at each iteration $k \geq 0$, we generate an estimator \tilde{S}_γ^k satisfying Definition 2.1 and update

$$x^{k+1} := x^k - \eta \tilde{S}_\gamma^k - \eta(\gamma v^{k+1} - (2\gamma - 1)v^k), \quad (\text{VFRBS})$$

where $\eta > 0$ and $\gamma > 0$ are determined later, $v^k \in Tx^k$, $x^{-1} = x^{-2} := x^0$, and $\tilde{S}_\gamma^0 := (1 - \gamma)Gx^0$.

(b) **Implementable version.** Since $v^{k+1} \in Tx^{k+1}$ appears on the RHS of (VFRBS), using the resolvent $J_{\gamma\eta T} := (\mathbb{I} + \gamma\eta T)^{-1}$ of T , we can rewrite (VFRBS) equivalently to

$$\begin{cases} y^{k+1} := x^k - \eta \tilde{S}_\gamma^k + \frac{(2\gamma-1)}{\gamma}(y^k - x^k), \\ x^{k+1} := J_{\gamma\eta T}(y^{k+1}). \end{cases} \quad (10)$$

Here, $y^0 \in \text{dom}(\Psi)$ is given, and $x^0 = x^{-1} := J_{\gamma\eta T}(y^0)$. This is an implementable variant of (VFRBS) using the resolvent $J_{\gamma\eta T}$. Clearly, if $\gamma = \frac{1}{2}$, then (10) reduces to

$$x^{k+1} := J_{(\eta/2)T}(x^k - \eta \tilde{S}_{1/2}^k),$$

which can be viewed as a stochastic forward-reflected-backward splitting scheme. However, our $\gamma \in (\frac{1}{2}, 1)$, making (10) different from existing methods, even in the deterministic case.

Compared to (Alacaoglu & Malitsky, 2022), (10) requires only one $J_{\gamma\eta T}$ as in (Alacaoglu et al., 2023), while (Alacaoglu & Malitsky, 2022) needs more than ones. Moreover, our estimator \tilde{S}_γ^k is also different from (Alacaoglu & Malitsky, 2022). Compared to (Beznosikov et al., 2023) and (Alacaoglu et al., 2023), the term $\gamma^{-1}(2\gamma - 1)(y^k - x^k)$ makes it different from SGDA in (Beznosikov et al., 2023) and the golden-ratio method in (Alacaoglu et al., 2023), and also other existing deterministic methods.

(c) **Approximate solution certification.** To certify an approximate solution of (NI), we note that its exact solution $x^* \in \text{zer}(\Psi)$ satisfies $\|Gx^* + v^*\|^2 = 0$ for some $v^* \in Tx^*$. Therefore, if (x^k, v^k) satisfies $\mathbb{E}[\|Gx^k + v^k\|^2] \leq \epsilon^2$ for some $v^k \in Tx^k$, then we can say that x^k is an ϵ -solution of (NI). Alternatively, we can define the following forward-backward splitting (FBS) residual of (NI):

$$\mathcal{G}_\eta x := \eta^{-1}(x - J_\eta(x - \eta Gx)),$$

for any given $\eta > 0$. It is well-known that $x^* \in \text{zer}(\Psi)$ iff $\mathcal{G}_\eta x^* = 0$. Hence, if $\mathbb{E}[\|\mathcal{G}_\eta x^k\|^2] \leq \epsilon^2$, then x^k is also called an ϵ -solution of (NI). One can easily prove that $\|\mathcal{G}_\eta x^k\| \leq \|Gx^k + v^k\|$ for any $v^k \in Tx^k$. Clearly, the former metric implies the latter one. Therefore, it is sufficient to only certify $\mathbb{E}[\|Gx^k + v^k\|^2] \leq \epsilon^2$, which implies $\mathbb{E}[\|\mathcal{G}_\eta x^k\|^2] \leq \epsilon^2$.

(d) **Convergence analysis.** For simplicity of our presentation, for a given $\gamma \in (\frac{1}{2}, 1)$, with ρ , C , and \hat{C} in Definition 2.1, we define the following two parameters:

$$M := 4\gamma^2 + \frac{4\gamma}{1-\gamma} \cdot \frac{C+\hat{C}}{\rho} \quad \text{and} \quad \delta := \frac{\gamma(2\gamma-1)}{(3\gamma-1)\sqrt{M}}. \quad (11)$$

Then, Theorem 4.1 below states the convergence of (VFRBS), whose proof can be found in Supp. Doc. D.

Theorem 4.1. Let us fix $\gamma \in (\frac{1}{2}, 1)$, and define M and δ as in (11). Suppose that Assumptions 1.1, 1.2, 1.3, and 1.4 hold for (NI) for some $\kappa \geq 0$ such that $L\kappa < \delta$. Let $\{x^k\}$ be generated by (VFRBS) using a fixed learning rate η such that $\frac{(3\gamma-1)\kappa}{\gamma(2\gamma-1)} < \eta \leq \frac{1}{L\sqrt{M}}$. Then, we have

$$\begin{aligned} \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k + v^k\|^2] &\leq \frac{\hat{\Theta}_1 \hat{R}_0^2}{\eta^2(K+1)}, \\ \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|x^k - x^{k-1}\|^2] &\leq \frac{\hat{\Theta}_2 \hat{R}_0^2}{K+1}, \end{aligned} \quad (12)$$

where \hat{R}_0^2 , $\hat{\Theta}_1$, and $\hat{\Theta}_2$ are respectively given by

$$\begin{aligned} \hat{R}_0^2 &:= \|x^0 - x^*\|^2 + \gamma^2 \eta^2 \|Gx^0 + v^0\|^2, \\ \hat{\Theta}_1 &:= \frac{(3\gamma-1)\eta}{(1-\gamma)[\gamma(2\gamma-1)\eta - (3\gamma-1)\kappa]}, \\ \hat{\Theta}_2 &:= \frac{4(3\gamma-1)\hat{R}_0^2}{(1-\gamma)(1-ML^2\eta^2)}. \end{aligned}$$

The bounds in Theorem 4.1 are similar to Theorem 3.1, but their proof relies on a new Lyapunov function. Note that the condition on $L\kappa$ still depends on ρ as $L\kappa \leq \delta = \mathcal{O}(\sqrt{\rho})$.

4.2. Complexity of VFRBS with SVRG and SAGA

Similar to Section 3, we can apply Theorem 4.1 for the mini-batch SVRG estimator in Section 2.

Corollary 4.2. Suppose that Assumptions 1.1, 1.2, 1.3, and 1.4 hold for (NI) with $\kappa \geq 0$ as in Theorem 4.1. Let $\{x^k\}$ be generated by (VFRBS) using the SVRG estimator (L-SVRG), $\gamma \in (\frac{1}{2}, 1)$, and $\eta := \frac{1}{L\sqrt{M}} \geq \frac{\sigma\sqrt{b}\mathbf{p}}{L}$ with $\sigma := \frac{\sqrt{1-\gamma}}{2\sqrt{8+\gamma+7\gamma^2}}$, provided that $b\mathbf{p}^2 \leq 1$. Then, we have

$$\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k + v^k\|^2] \leq \frac{\hat{\Theta}_1 L^2 \hat{R}_0^2}{\sigma^2 b \mathbf{p}^2 (K+1)}, \quad (13)$$

where $\hat{R}_0^2 := \|x^0 - x^*\|^2 + \gamma^2 \eta^2 \|Gx^0 + v^0\|^2$.

For given $\epsilon > 0$, if we choose $\mathbf{p} := n^{-1/3}$ and $b := \lfloor n^{2/3} \rfloor$, then (VFRBS) requires $\mathcal{T}_{G_i} := n + \lfloor \frac{4\Gamma L^2 \hat{R}_0^2 n^{2/3}}{\epsilon^2} \rfloor$ evaluations of G_i and $\mathcal{T}_T = \lfloor \frac{\Gamma L^2 \hat{R}_0^2}{\epsilon^2} \rfloor$ evaluations of $J_{\gamma\eta T}$ to achieve $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k + v^k\|^2] \leq \epsilon^2$, where $\Gamma := \frac{\hat{\Theta}_1}{\sigma^2}$.

Alternatively, we can apply Theorem 4.1 to the mini-batch SAGA estimator (SAGA) in Section 2.

Corollary 4.3. Suppose that Assumptions 1.1, 1.2, 1.3, and 1.4 hold for (NI) with $\kappa \geq 0$ as in Theorem 4.1. Let $\{x^k\}$ be generated by (VFRBS) using the SAGA estimator (SAGA), $\gamma \in (\frac{1}{2}, 1)$, and $\eta := \frac{1}{L\sqrt{M}} \geq \frac{\sigma b^{3/2}}{nL}$ with $\sigma := \frac{\sqrt{1-\gamma}}{2\sqrt{\gamma(10+\gamma+7\gamma^2)}}$, provided that $1 \leq b \leq n^{2/3}$. Then

$$\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k + v^k\|^2] \leq \frac{n^2 \hat{\Theta}_1 L^2 \hat{R}_0^2}{\sigma^2 b^3 (K+1)}, \quad (14)$$

where $\hat{R}_0^2 := \|x^0 - x^*\|^2 + \gamma^2 \eta^2 \|Gx^0 + v^0\|^2$.

For a given $\epsilon > 0$, if we choose $b := \lfloor n^{2/3} \rfloor$, then (VFRBS) requires $\mathcal{T}_{G_i} := n + \lfloor \frac{3\Gamma L^2 \hat{R}_0^2 n^{2/3}}{\epsilon^2} \rfloor$ evaluations of G_i and $\mathcal{T}_T := \lfloor \frac{\Gamma L^2 \hat{R}_0^2}{\epsilon^2} \rfloor$ evaluations of $J_{\gamma\eta T}$ to achieve $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k + v^k\|^2] \leq \epsilon^2$, where $\Gamma := \frac{\hat{\Theta}_1}{\sigma^2}$.

Similar to Subsection 3.2, when γ , n , b , and \mathbf{p} are given, we can compute concrete values of the theoretical learning rate η in both corollaries. They are larger than the corresponding lower bounds given in these corollaries.

5. Numerical Experiments

We provide two examples to illustrate (VFR) and (VFRBS) and compare them with other methods.

5.1. Nonconvex-nonconcave minimax optimization

We consider the following nonconvex-nonconcave minimax optimization problem as a special case of (SP):

$$\min_{u \in \mathbb{R}^{p_1}} \max_{v \in \mathbb{R}^{p_2}} \{\mathcal{L}(u, v) := \varphi(u) + \mathcal{H}(u, v) - \psi(v)\}, \quad (15)$$

where $\mathcal{H}(u, v) := \frac{1}{n} \sum_{i=1}^n \mathcal{H}_i(u, v) = \frac{1}{n} \sum_{i=1}^n [u^T A_i u + u^T L_i v - v^T B_i v + b_i^T u - c_i^T v]$ such that $A_i \in \mathbb{R}^{p_1 \times p_1}$ and $B_i \in \mathbb{R}^{p_2 \times p_2}$ are symmetric matrices, $L_i \in \mathbb{R}^{p_1 \times p_2}$, $b_i \in \mathbb{R}^{p_1}$, and $c_i \in \mathbb{R}^{p_2}$; φ and ψ are two proper, closed, and convex functions. The optimality of (15) becomes (NI) (see Supp. Doc. E). In our experiments, we choose A_i and B_i to be not positive semidefinite such that Assumption 1.4 holds. Thus, (15) is nonconvex-nonconcave.

We generate $A_i = Q_i D_i Q_i^T$ for a given orthonormal matrix Q_i and a diagonal matrix D_i , where its entries D_i^j are generated from standard normal distribution and clipped by $\max\{D_i^j, -0.1\}$. The matrix B_i is also generated by the same way, while L_i , b_i , and c_i are generated from standard normal distribution. Hence, \mathbf{G} in (NI) is not symmetric and also not positive semidefinite.

The unconstrained case. We implement three variants of (VFR): VFR-svrg (double-loop SVRG), LVFR-svrg (loopless SVRG), VFR-saga (using SAGA estimator) in Python to solve (15) when both φ and ψ are vanished, i.e. its optimality is a special case of (NE). We also compare our methods with the deterministic optimistic gradient method (OG) in (Daskalakis et al., 2018), the variance-reduced

FRBS scheme (VFRBS) in (Alacaoglu et al., 2023), and the variance-reduced extragradient algorithm (VEG) in (Alacaoglu & Malitsky, 2022). We select the parameters as suggested by our theory, while choosing appropriate parameters for OG, VFRBS, and VEG. The details of this experiment, including generating data and specific choice of parameters, are given in Supp. Doc. E.

The relative residual norm $\|Gx^k\|/\|Gx^0\|$ against the number of epochs averaged on 10 problem instances is revealed in Figure 1 for two datasets $(p, n) = (100, 5000)$ and $(p, n) = (200, 10000)$.

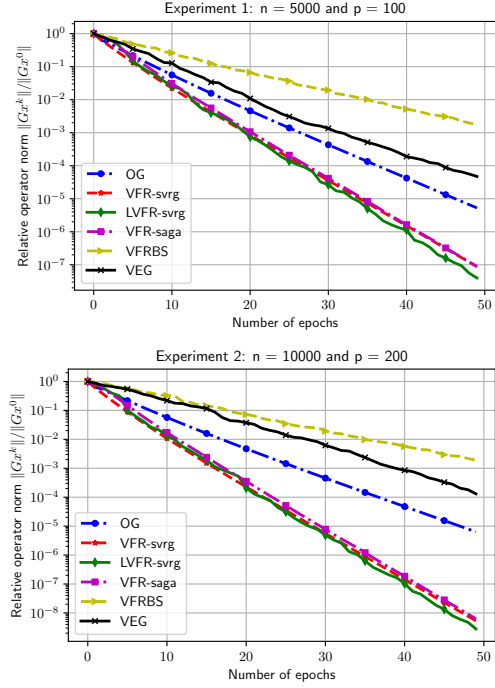


Figure 1: Comparison of 6 algorithms to solve the unconstrained (15) on 2 experiments (The average of 10 runs).

Clearly, with these experiments, three SVRG variants of our method (VFRBS) work well and significantly outperform other competitors. The LVFR-svrg variant of (VFRBS) seems to work best, while VFRBS and VEG still cannot beat the deterministic algorithm OG in this example.

The constrained case. We now adding two simplex constraints $u \in \Delta_{p_1}$ and $v \in \Delta_{p_2}$ to (15), where $\Delta_p := \{u \in \mathbb{R}_+^p : \sum_{i=1}^p u_i = 1\}$ is the standard simplex in \mathbb{R}^p . These constraints are common in bilinear games. To handle these constraints, we set $\varphi(u) := \delta_{\Delta_{p_1}}(u)$ and $\psi(v) := \delta_{\Delta_{p_2}}(v)$ as the indicators of Δ_{p_1} and Δ_{p_2} , respectively.

Again, we run 6 algorithms for solving this constrained case of (15) using the same parameters as **the unconstrained case**. We report the relative norm of the FBS residual $\|\mathcal{G}_\eta x^k\|/\|\mathcal{G}_\eta x^0\|$ against the number of epochs. The results are revealed in Figure 2 for two datasets $(p, n) = (100, 5000)$ and $(p, n) = (200, 10000)$.

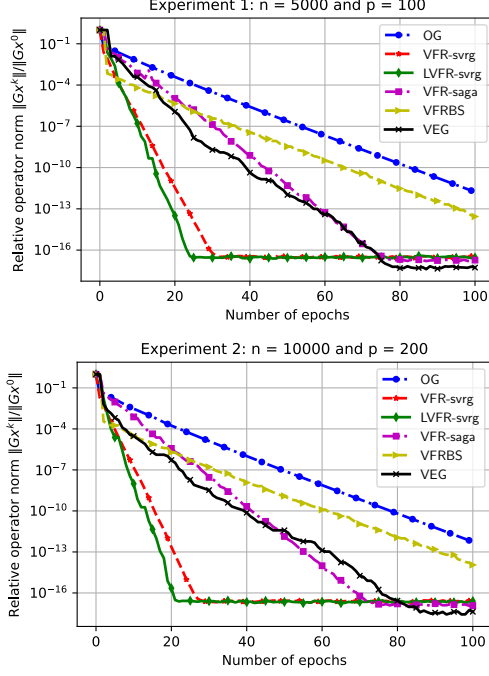


Figure 2: The performance of 6 algorithms to solve the constrained (15) on 2 experiments (The average of 10 runs).

Clearly, with these experiments, both SVRG variants of our method (VFRBS) work well and significantly outperform other competitors. The SVRG variant (VFR-svrg) of (VFRBS) seems to work best, while our VFR-saga has a similar performance as VEG. Again, we also see that VFRBS tends to have a similar performance as OG.

5.2. Logistic regression with ambiguous features

We consider the following minimax optimization problem arising from a regularized logistic regression with ambiguous features (see Supp. Doc. E for more details):

$$\min_{w \in \mathbb{R}^d} \max_{z \in \mathbb{R}^m} \left\{ \mathcal{L}(w, z) := \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m z_j \ell(\langle X_{ij}, w \rangle, y_i) + \tau R(w) - \delta_{\Delta_m}(z) \right\}, \quad (16)$$

where $\ell(\tau, s) := \log(1 + \exp(\tau)) - s\tau$ is the standard logistic loss, $R(w) := \|w\|_1$ is an ℓ_1 -norm regularizer, $\tau > 0$ is a regularization parameter, and δ_{Δ_m} is the indicator of Δ_m to handle the constraint $z \in \Delta_m$. Then, the optimality condition of (16) can be cast into (NI), where $x := [w, z]$.

We implement three variants of (VFRBS) to solve (16): VFR-svrg, LVFR-svrg, and VFR-saga. We also compare our methods with OG, VFRBS, and VEG as in Subsection 5.1. We carry out a manual tuning procedure to select appropriate learning rates for all methods. We test these algorithms on two real datasets: a9a (134 features and 3561 samples) and w8a (311 features and 45546 samples) downloaded from LIBSVM (Chang & Lin, 2011). We first normalize the feature vector \hat{X}_i and add a column of all ones to address the bias term. To generate ambiguous features,

we take the nominal feature vector \hat{X}_i and add a random noise generated from a normal distribution of zero mean and variance of $\sigma^2 = 0.5$. In our test, we choose $\tau := 10^{-3}$ and $m := 10$. The relative FBS residual norm $\|\mathcal{G}_\eta x^k\|/\|\mathcal{G}_\eta x^0\|$ against the epochs is plotted in Figure 3 for both datasets.

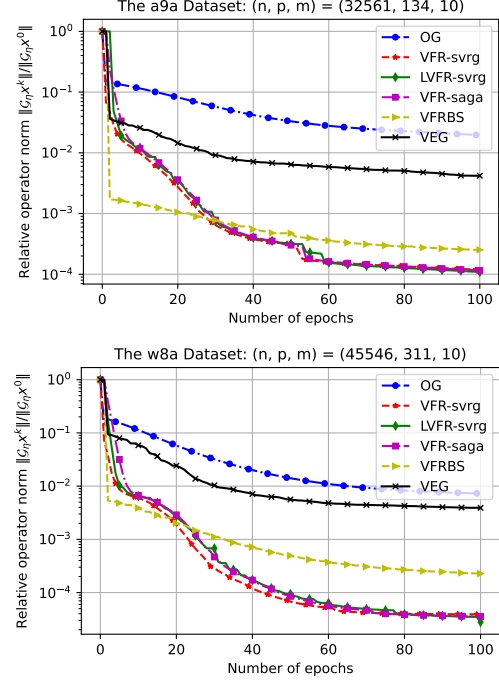


Figure 3: Comparison of 6 algorithms to solve (16) on two real datasets: a9a and w8a.

As we can observe from Figure 3 that three variants VFR-svrg, LVFR-svrg, and VFR-saga have similar performance and are better than their competitors. Among three competitors, VFRBS still works well, and is much better than OG and VEG. The deterministic method, OG, is the worst one in terms of oracle complexity. In this test, VEG has a larger learning rate than ours and VFRBS.

6. Conclusions

We develop two new variance-reduced algorithms based on the forward-reflected-backward splitting method to tackle both root-finding problems (NE) and (NI). These methods encompass both SVRG and SAGA estimators as special cases. By carefully selecting the parameters, our algorithms achieve the state-of-the-art oracle complexity for attaining an ϵ -solution, matching the state-of-the-art complexity bounds observed in nonconvex optimization methods using SVRG and SAGA. While the first scheme resembles a stochastic variant of the optimistic gradient method, the second one is entirely novel and distinct from existing approaches, even their deterministic counterparts. We have validated our methods through numerical examples, and the results demonstrate promising performance compared to existing techniques under careful parameter selections.

Impact Statement

This paper proposes new algorithms with rigorous convergence guarantees and complexity estimates for solving a broad class of large-scale problems. These problems cover many fundamental challenges and applications in optimization, machine learning, and related fields as special cases. We believe that our new algorithms have the potential to make a significant impact in machine learning and related areas. Additionally, there are various potential societal consequences of our work, though none that we feel require specific emphasis at this time.

Acknowledgements

This work was partly supported by the National Science Foundation (NSF): NSF-RTG grant No. NSF DMS-2134107 and the Office of Naval Research (ONR), grant No. N00014-23-1-2588.

References

- Alacaoglu, A. and Malitsky, Y. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 778–816. PMLR, 2022.
- Alacaoglu, A., Malitsky, Y., and Cevher, V. Forward-reflected-backward method with variance reduction. *Comput. Optim. Appl.*, 80(2):321–346, 2021.
- Alacaoglu, A., Böhm, A., and Malitsky, Y. Beyond the golden ratio for variational inequality algorithms. *J. Mach. Learn. Res.*, 24(172):1–33, 2023.
- Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pp. 699–707. PMLR, 2016.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.
- Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.
- Bauschke, H. H. and Combettes, P. *Convex analysis and monotone operators theory in Hilbert spaces*. Springer-Verlag, 2nd edition, 2017.
- Bauschke, H. H., Moursi, W. M., and Wang, X. Generalized monotone operators and their averaged resolvents. *Math. Program.*, pp. 1–20, 2020.
- Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*. Princeton University Press, 2009.
- Bertsimas, D. Brown, D. and Caramanis, C. Theory and Applications of Robust Optimization. *SIAM Review*, 53(3):464–501, 2011.
- Beznosikov, A., Gorbunov, E., Berard, H., and Loizou, N. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 172–235. PMLR, 2023.
- Bhatia, K. and Sridharan, K. Online learning with dynamics: A minimax perspective. *Advances in Neural Information Processing Systems*, 33:15020–15030, 2020.
- Böhm, A. Solving nonconvex-nonconcave min-max problems exhibiting weak Minty solutions. *Transactions on Machine Learning Research*, 2022.
- Bot, R. I., Mertikopoulos, P., Staudigl, M., and Vuong, P. T. Forward-backward-forward methods with variance reduction for stochastic variational inequalities. *arXiv preprint arXiv:1902.03355*, 2019.
- Burachik, R. S. and Iusem, A. *Set-Valued Mappings and Enlargements of Monotone Operators*. New York: Springer, 2008.
- Cai, X., Song, C., Guzmán, C., and Diakonikolas, J. A stochastic Halpern iteration with variance reduction for stochastic monotone inclusion problems. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2022)*, 2022. URL <https://openreview.net/forum?id=BRZos-8TpCf>.
- Cai, X., Alacaoglu, A., and Diakonikolas, J. Variance reduced halpern iteration for finite-sum monotone inclusions. In *The 12th International Conference on Learning Representations (ICLR)*, pp. 1–33, 2024.
- Cai, Y. and Zheng, W. 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.
- Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. *Advances in Neural Information Processing Systems*, 32, 2019.
- Cevher, V. and Vü, B. A reflected forward-backward splitting method for monotone inclusions involving Lipschitzian operators. *Set-Valued and Variational Analysis*, 29(1):163–174, 2021.
- Chang, C.-C. and Lin, C.-J. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

- Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32:393–403, 2019.
- Combettes, P. L. and Pennanen, T. Proximal methods for cophyomontone operators. *SIAM J. Control Optim.*, 43 (2):731–742, 2004.
- Cui, S. and Shanbhag, U. 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.
- Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, pp. 15210–15219, 2019.
- Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with Optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.
- Davis, D. Variance reduction for root-finding problems. *Math. Program.*, pp. 1–36, 2022.
- Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1646–1654, 2014.
- Diakonikolas, J. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pp. 1428–1451. PMLR, 2020.
- Diakonikolas, J., Daskalakis, C., and Jordan, M. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.
- Driggs, D., Ehrhardt, M. J., and Schönlieb, C.-B. Accelerating variance-reduced stochastic gradient methods. *Math. Program.*, (online first):1–45, 2020.
- Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*, volume 1-2. Springer-Verlag, 2003.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Gorbunov, E., Berard, H., Gidel, G., and Loizou, N. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pp. 7865–7901. PMLR, 2022.
- Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Math. Program.*, 188(1):135–192, 2021.
- Hanzely, F., Mishchenko, K., and Richtárik, P. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pp. 2082–2093, 2018.
- Huang, K., Wang, N., and Zhang, S. An accelerated variance reduced extra-point approach to finite-sum vi and optimization. *arXiv preprint arXiv:2211.03269*, 2022.
- Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.*, 27(2):686–724, 2017.
- Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.
- Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- Kannan, A. and Shanbhag, U. V. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Comput. Optim. Appl.*, 74(3):779–820, 2019.
- Kotsalis, G., Lan, G., and Li, T. Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation. *SIAM J. Optim.*, 32(3):2041–2073, 2022.
- Kovalev, D., Horvath, S., and Richtarik, P. Don’t jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pp. 451–467. PMLR, 2020.
- Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS2021)*, 2021.
- Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.
- Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. 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.

- Luo, Y. and Tran-Dinh, Q. Extragradient-type methods for co-monotone root-finding problems. (*UNC-STOR Technical Report*), 2022.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- Malitsky, Y. Projected reflected gradient methods for monotone variational inequalities. *SIAM J. Optim.*, 25(1):502–520, 2015.
- Malitsky, Y. and Tam, M. K. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM J. Optim.*, 30(2):1451–1472, 2020.
- Namkoong, H. and Duchi, J. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.
- Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 109(2–3):319–344, 2007.
- Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2613–2621, 2017.
- Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pp. 1416–1424, 2016.
- Pethick, T., Patrinos, P., Fercoq, O., and Cevher, V. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.
- Pethick, T., Fercoq, O., Latafat, P., Patrinos, P., and Cevher, V. Solving stochastic weak Minty variational inequalities without increasing batch size. In *Proceedings of International Conference on Learning Representations (ICLR)*, pp. 1–34, 2023.
- Phelps, R. R. *Convex functions, monotone operators and differentiability*, volume 1364. Springer, 2009.
- Popov, L. D. A modification of the Arrow-Hurwicz method for search of saddle points. *Math. notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.
- Reddi, S. J., Hefny, A., Sra, S., Póczos, B., and Smola, A. J. Stochastic variance reduction for nonconvex optimization. In *ICML*, pp. 314–323, 2016a.
- Reddi, S. J., Sra, S., Póczos, B., and Smola, A. Fast incremental method for smooth nonconvex optimization. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 1971–1977. IEEE, 2016b.
- Robbins, H. and Monroe, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.
- Rockafellar, R. and Wets, R. J.-B. *Variational Analysis*. Springer-Verlag, 1997.
- Ryu, E. and Yin, W. *Large-scale convex optimization: Algorithms & analyses via monotone operators*. Cambridge University Press, 2022.
- Ryu, E. K. and Boyd, S. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1):3–43, 2016.
- Tran-Dinh, Q. Extragradient-Type Methods with $\mathcal{O}(1/k)$ -Convergence Rates for Co-Hypomonotone Inclusions. *J. Global Optim.*, pp. 1–25, 2023a.
- Tran-Dinh, Q. Sublinear Convergence Rates of Extragradient-Type Methods: A Survey on Classical and Recent Developments. *arXiv preprint arXiv:2303.17192*, 2023b.
- Tran-Dinh, Q. Accelerated Variance-Reduced Forward-Reflected Methods for Root-Finding Problems. *arXiv preprint arXiv:2406.02413*, 2024.
- Tran-Dinh, Q. and Luo, Y. Randomized Block-Coordinate Optimistic Gradient Algorithms for Root-Finding Problems. *Math. Oper. Res.*, in press, 2025.
- Tran-Dinh, Q., Pham, H. N., Phan, T. D., and Nguyen, M. L. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *Preprint: arXiv:1905.05920*, 2019.
- Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. *Math. Program.*, 191: 1005–1071, 2022.
- Yousefian, F., Nedić, A., and Shanbhag, U. V. 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.
- Yu, Y., Lin, T., Mazumdar, E. V., and Jordan, M. Fast distributionally robust learning with variance-reduced minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1219–1250. PMLR, 2022.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3925–3936. Curran Associates Inc., 2018.

Supplementary Document:

Variance-Reduced Forward-Reflected-Backward Splitting Methods for Nonmonotone Generalized Equations

Due to space limit, some parts of our algorithmic construction and theory are not described in detail and motivated in the main text. This supplementary document aims at providing more details of the algorithmic construction, motivation, related work, technical proofs, and additional experiments related to our methods.

A. Further Discussion of Related Work and Assumptions

Let us further expand our discussion of related work in the main text. Then, we show that our assumptions, Assumptions 1.3 and 1.4, indeed cover nonmonotone problems.

A.1. Further Discussion of Related Work

As we already discussed in the introduction of the main text, both standard stochastic approximation and variance-reduction methods have been broadly studied for (NE) and (NI), including (Juditsky et al., 2011; Kotsalis et al., 2022; Pethick et al., 2023). In this section, we further discuss some other related work to (NE) and (NI), their special cases, and equivalent forms.

(a) **Beyond monotonicity.** Classical methods such as extragradient, prox-mirror, and projective schemes often relax the monotonicity to star-monotonicity, and other forms such as pseudo-monotonicity and quasi-monotonicity (Konnov, 2001; Noor, 2003; Noor & Al-Said, 1999; Tu, 2018). These assumptions are certainly weaker than the monotonicity and can cover some wider classes of problems, including some nonmonotone subclasses. Another extension of monotonicity is the weak-Minty solution condition in Assumption 1.4, which was proposed in early work, perhaps in the most recent one such as (Diakonikolas et al., 2021), as an extension of the star-monotonicity and star-weak-monotonicity assumptions. Other following-up works include (Böhm, 2022; Gorbunov et al., 2022b; Luo & Tran-Dinh, 2022). A comprehensive survey for extragradient-type methods using the weak-Minty solution condition can be found in (Tran-Dinh, 2023; Tran-Dinh & Nguyen-Trung, 2025a;b). The monotonicity has also been extended to a weak monotonicity, or related, prox-regularity (Rockafellar & Wets, 1997) (in particular, weak-convexity). Other types of hypo-monotonicity or co-monotonicity concepts can be found, e.g., in (Bauschke et al., 2020). These concepts have been exploited to develop algorithms for solving (NE) and (NI) and their special cases. For stochastic methods, extensions beyond monotonicity have been also extensively explored. For instance, some further structures beyond monotonicity such as weak solution were exploited for MVIs in (Song et al., 2020), a pseudo-monotonicity was used in (Boş et al., 2021; Kannan & Shanbhag, 2019) for stochastic VIPs, a two-sided Polyak-Łojasiewicz condition was extended to VIP in (Yang et al., 2020) to tackle a class on nonconvex-nonconcave minimax problems, an expected co-coercivity was used (Loizou et al., 2021), and a strongly star-monotone was further exploited in (Gorbunov et al., 2022a). While these structures are occasionally used in different works, the relation between them is still largely elusive. In addition, their relation to concrete applications is still not well studied.

(b) **Further discussion on stochastic methods.** Under the monotonicity, several authors have exploited the stochastic approximation approach (Robbins & Monro, 1951) to develop stochastic variants for solving (NE) and (NI) and their special cases. For example, a stochastic Mirror-Prox was proposed in (Juditsky et al., 2011), which has convergence on a gap function, but requires a bounded domain assumption. This approach was later extended to the extragradient method under additional assumptions in (Mishchenko et al., 2020). In (Hsieh et al., 2019), the authors discussed several methods for solving MVIs, a special case of (NI), including stochastic methods. They experimented on numerical examples and showed that the norm of the operator can asymptotically converge for unconstrained MVIs with a double learning rate. In the last few years, there were many works focusing on developing stochastic methods for solving (NE) and (NI), and their special cases using different techniques such as single-call stochastic schemes in (Hsieh et al., 2019), non-accelerated and accelerated variance reduction with Halpern-type iterations in (Cai et al., 2024; 2022), co-coercive structures in (Beznosikov et al., 2023), and bilinear game models in (Li et al., 2022).

(c) **The challenge of using biased estimators in algorithms for solving (NI) and related problems.** In optimization, especially in nonconvex optimization, stochastic methods using biased estimators such as SARAH, Nested SVRG, Hybrid-SGD, and STORM can achieve better, even “optimal” oracle complexity compared to unbiased ones such as standard SVRG and SAGA, see, e.g., (Cutkosky & Orabona, 2019; Driggs et al., 2022; Pham et al., 2020; Tran-Dinh et al., 2022). However,

it becomes challenging in root-finding algorithms for solving (NI) and related problems such as minimax optimization and VIPs. One main reason for this is that the convergence analysis of these optimization methods relies on the objective function as a key metric to prove convergence guarantee and to estimate oracle complexity. However, in (NI) and (NE), we do not have such an object, making it difficult to process the biased terms, including product terms such as $\langle e^k, x^{k+1} - x^* \rangle$ and $\langle e^k, x^{k+1} - x^k \rangle$, where e^k is a bias rendered from the underlying stochastic estimator. This is currently one of the main obstacles to move from using unbiased to biased estimators in root-finding algorithms, including our methods in this paper.

(d) **Comparison to (Cai et al., 2024).** Among many existing works, perhaps, (Cai et al., 2024) is one of the most recent works that develops variance-reduction methods for solving (NI) and achieves the state-of-the-art oracle complexity. However, (Cai et al., 2024) explores a different approach than ours, which relies on some recent development of the Halpern fixed-point iteration and a biased SARAH estimator. Let us clarify the differences of this work and our paper here. Algorithm 1 in (Cai et al., 2024) is a single-loop and achieves a better oracle complexity. However, it requires a much stronger assumption, Assumption 3, which is a co-coercive condition. Note that this assumption excludes the well-known bilinear matrix game, or the synthetic WGAN model (37) below. Section 4 of (Cai et al., 2024) studies both the monotone and the co-hypomonotone cases of (NI). The main idea is to reformulate (NI) into a resolvent equation $J_{\eta(G+T)}x = 0$ and then apply a deterministic variant of Algorithm 1 to this equation, where $J_{\eta(G+T)}$ is co-coercive. However, exactly evaluating $J_{\eta(G+T)}$ is impractical, one needs to approximate it by an appropriate algorithm. For instance, (Cai et al., 2024) suggests to use the variance-reduced FRBS method in (Alacaoglu et al., 2023) to approximate this resolvent, leading to a double loop algorithm. Note that this method also relies on a unbiased estimator, namely SVRG. This approach is not a direct variance-reduced method (i.e., the inner loop can be any algorithm) as ours or Algorithm 1 of (Cai et al., 2024). Moreover, practically implementing as well as rigorously analyzing an inexact double loop algorithm, when the inner loop is also a stochastic method, is often very challenging and technical as it is difficult to conduct a stopping criterion of the inner loop, and to select appropriate parameters. Nevertheless, our algorithms developed in this paper are simple to implement and applicable to both (NE) and (NI) whose weak-Minty solution exists. These problems are broader than the ones in (Cai et al., 2024). We also believe that our oracle complexity in this paper can be further improved by exploiting enhancement techniques such as nested trick or multiple loops as done in (Cai et al., 2024; Zhou et al., 2018).

(e) **Randomized coordinate and cyclic coordinate methods for (NE) and (NI).** Together with stochastic algorithms for solving (NE) and (NI) and their special cases, randomized coordinate methods have also been proposed to solve these problems, including (Combettes & Eckstein, 2018; Combettes & Pesquet, 2015; Peng et al., 2016). Recent works on randomized coordinate and cyclic coordinate methods can be found, e.g., in (Chakrabarti et al., 2024; Cui & Shanbhag, 2021; Hamedani et al., 2018; Song & Diakonikolas, 2023; Tran-Dinh & Luo, 2025; Yousefian et al., 2018). These methods are not directly related to our work, but they can be considered as a dual form of stochastic methods in certain settings such as convex-concave minimax problems. Studying relations between randomized coordinate methods and stochastic algorithms for (NE) and (NI) appears to be an interesting research topic.

A.2. A Nonmonotone Example (NI)

As an example of (NI), we can consider the following linear operators

$$Gx := \mathbf{G}x + \mathbf{g} \quad \text{and} \quad Tx := \mathbf{T}x,$$

where \mathbf{G} and \mathbf{T} are given square matrices in $\mathbb{R}^{p \times p}$ and $\mathbf{g} \in \mathbb{R}^p$ is a given vector. Clearly, for any \mathbf{G} , G is L -Lipschitz continuous with $L := \|\mathbf{G}\|$ (the operator norm of \mathbf{G}). Our goal is to choose \mathbf{G} and \mathbf{T} such that $\Psi x := Gx + Tx$ is nonmonotone and satisfies Assumption 1.4.

- Clearly, we can choose \mathbf{G} and \mathbf{T} such that $\frac{1}{2}(\mathbf{G} + \mathbf{G}^\top)$ is positive semidefinite and $\frac{1}{2}(\mathbf{G} + \mathbf{G}^\top + \mathbf{T} + \mathbf{T}^\top)$ is not positive semidefinite. This shows that Ψ is nonmonotone.
- Now, Assumption 1.4 holds if Ψ is κ -co-hypomonotone, i.e. there exists $\kappa > 0$ such that $\langle u - v, x - y \rangle \geq -\kappa \|u - v\|^2$, for any $(x, u), (y, v) \in \text{gra}(\Psi)$. In the linear case, this condition is equivalent to

$$\mathbf{S} := \frac{1}{2}(\mathbf{G} + \mathbf{G}^\top + \mathbf{T} + \mathbf{T}^\top) + \kappa(\mathbf{G} + \mathbf{T})^\top(\mathbf{G} + \mathbf{T}) \quad \text{is positive semidefinite.}$$

However, since $\frac{1}{2}(\mathbf{G} + \mathbf{G}^\top)$ is positive semidefinite, this condition holds if $\frac{1}{2}(\mathbf{T} + \mathbf{T}^\top) + \kappa(\mathbf{G} + \mathbf{T})^\top(\mathbf{G} + \mathbf{T})$ is positive semidefinite. In particular, Ψ satisfies Assumption 1.4.

For example, given any $\epsilon > 0$, we choose

$$\mathbf{G} := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{T} := \begin{bmatrix} -\epsilon & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad \kappa := \epsilon > 0.$$

Then, it is clear that \mathbf{G} is symmetric and positive semidefinite, but $\mathbf{G} + \mathbf{T}$ is symmetric and not positive semidefinite. Thus, $\Psi x = Gx + Tx$ is nonmonotone. Moreover, $Gx = \mathbf{G}x + \mathbf{g}$ is L -Lipschitz continuous with $L = 1$.

Next, we check Assumption 1.4. Clearly, \mathbf{G} is symmetric and positive semidefinite. Moreover, since \mathbf{G} and \mathbf{T} are symmetric, we have

$$\mathbf{M} := \frac{1}{2}(\mathbf{T} + \mathbf{T}^\top) + \kappa(\mathbf{G} + \mathbf{T})^\top(\mathbf{G} + \mathbf{T}) = \begin{bmatrix} -\epsilon & 0 \\ 0 & 0 \end{bmatrix} + \kappa \begin{bmatrix} -\epsilon & 1 \\ 1 & 0 \end{bmatrix}^\top \begin{bmatrix} -\epsilon & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -\epsilon + \kappa(1 + \epsilon^2) & -\kappa\epsilon \\ -\kappa\epsilon & \kappa \end{bmatrix}.$$

If we choose $\kappa = \epsilon$, then one can easily check that \mathbf{M} is positive semidefinite. Hence, we can conclude that Ψ is κ -cohypomonotone with $\kappa = \epsilon > 0$. In particular, Ψ satisfies Assumption 1.4. In addition, we can choose ϵ sufficiently small such that $L\kappa$ is sufficiently small, which fulfills the condition $L\kappa \leq \delta$ in Theorem 4.1.

B. The Proof of Technical Results in Section 2

This supplementary section provides the full proof of Lemma 2.2 and Lemma 2.3.

Further discussion of the FR quantity. Let us recall our quantity S_γ^k defined by (FRQ) as follows:

$$S_\gamma^k := Gx^k - \gamma Gx^{k-1}. \quad (\text{FRO})$$

As we mentioned earlier, γ plays a crucial role in our methods as $\gamma \in (\frac{1}{2}, 1)$. If $\gamma = \frac{1}{2}$, then we can write $S_{1/2}^k = \frac{1}{2}Gx^k + \frac{1}{2}(Gx^k - Gx^{k-1}) = \frac{1}{2}[2Gx^k - Gx^{k-1}]$ used in both the forward-reflected-backward splitting (FRBS) method (Malitsky & Tam, 2020) and the optimistic gradient method (Daskalakis et al., 2018).

Note that if we write $Gx^k - Gx^{k-1} = \hat{J}_G(x^k)(x^k - x^{k-1})$ by the Mean-Value Theorem, where $\hat{J}_G(x^k) := \int_0^1 \nabla G(x^{k-1} + \tau(x^k - x^{k-1}))d\tau$, then $S_\gamma^k = (1 - \gamma)G(x^k) + \gamma\hat{J}_G(x^k)(x^k - x^{k-1})$. Clearly, if γ is small, then S_γ^k can be considered as an approximation of Gx^k augmented by a second-order correction term $\gamma\hat{J}_G(x^k)(x^k - x^{k-1})$ (called Hessian-driven damping term or second-order dissipative term) widely used in dynamical systems for convex optimization, see, e.g., (Adly & Attouch, 2021; Attouch & Cabot, 2020). These two viewpoints motivate the use of our new operator S_γ^k , not only in our (VFR) and (VFRBS), but in other methods such as accelerated algorithms. Thus, the results in Section 2 are of independent interest, and can potentially be used to develop other methods.

Other possible stochastic estimators for S_γ^k . One natural idea to construct an unbiased estimator for S^k is to use an increasing mini-batch stochastic estimator as $\tilde{S}_\gamma^k := \frac{1}{b_k} \sum_{i \in \mathcal{B}_k} [G_i x^k - \gamma G_i x^{k-1}]$, where \mathcal{B}_k is an increasing mini-batch in $[n]$, with $b_k := |\mathcal{B}_k| \geq \frac{b_{k-1}}{1-\rho_k} \geq b_{k-1}$, see, e.g., (Iusem et al., 2017). While this idea may work well for the general expectation case $Gx = \mathbb{E}_\xi [G(x, \xi)]$, it may not be an ideal choice for the finite-sum operator (1) as $b_k \leq n$, which requires to stop increasing after finite iterations (i.e. $\mathcal{O}\left(\frac{\ln(n)}{-\ln(1-\rho)}\right)$ iterations). Other stochastic approximations may also fall into our class in Definition 2.1 such as JacSketch (Gower et al., 2021), SEGA (Hanzely et al., 2018), and quantized and compressed estimators (see, e.g., (Horváth et al., 2023)).

B.1. Proof of Lemma 2.2: Loopless-SVRG Estimator

Let us further expand Lemma 2.2 in detail as follows and then provide its full proof.

Lemma B.1. Let $S_\gamma^k := Gx^k - \gamma Gx^{k-1}$ be defined by (FRQ) and \tilde{S}_γ^k be generated by (L-SVRG). We define

$$\Delta_k := \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1 - \gamma)G_i w^k\|^2]. \quad (17)$$

Then, we have

$$\begin{aligned} \mathbb{E}_k[\tilde{S}_\gamma^k] &= S_\gamma^k \equiv Gx^k - \gamma Gx^{k-1}, \\ \mathbb{E}[\|\tilde{S}_\gamma^k - S_\gamma^k\|^2] &\leq \Delta_k - \frac{1}{b} \mathbb{E}[\|Gx^k - \gamma Gx^{k-1} - (1 - \gamma)Gw^k\|^2] \leq \Delta_k, \\ \Delta_k &\leq (1 - \frac{\mathbf{p}}{2})\Delta_{k-1} + \frac{(4-6\mathbf{p}+3\mathbf{p}^2)}{nb\mathbf{p}} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - G_i x^{k-1}\|^2] \\ &\quad + \frac{2\gamma^2(2-3\mathbf{p}+\mathbf{p}^2)}{nb\mathbf{p}} \sum_{i=1}^n \mathbb{E}[\|G_i x^{k-1} - G_i x^{k-2}\|^2]. \end{aligned} \quad (18)$$

Consequently, the SVRG estimator \tilde{S}_γ^k constructed by (L-SVRG) satisfies Definition 2.1 with Δ_k in (17), $\rho := \frac{\mathbf{p}}{2} \in (0, 1]$, $C := \frac{4-6\mathbf{p}+3\mathbf{p}^2}{b\mathbf{p}}$, and $\hat{C} := \frac{4\gamma^2(2-3\mathbf{p}+\mathbf{p}^2)}{b\mathbf{p}}$.

Proof. It is well-known, see, e.g., (Johnson & Zhang, 2013), that \tilde{S}_γ^k is an unbiased estimator of S^k conditioned on \mathcal{F}_k , we have $\mathbb{E}_k[\tilde{S}_\gamma^k] = S_\gamma^k$.

Next, let $X_i := G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)G_i w^k$ for any $i \in [n]$. Then, we have $\mathbb{E}_k[X_i] = Gx^k - \gamma Gx^{k-1} - (1-\gamma)Gw^k$ for any $i \in [n]$. Since \mathcal{B}_k is in \mathcal{F}_k , using the property of expectation, we can derive

$$\begin{aligned} \mathbb{E}_k[\|\tilde{S}_\gamma^k - S_\gamma^k\|^2] &\stackrel{(\text{L-SVRG})}{=} \mathbb{E}_k[\|\frac{1}{b} \sum_{i \in \mathcal{B}_k} X_i - [Gx^k + \gamma Gx^{k-1} - (1-\gamma)Gw^k]\|^2] \\ &= \mathbb{E}_k[\|\frac{1}{b} \sum_{i \in \mathcal{B}_k} [X_i - \mathbb{E}_k[X_i]]\|^2] \\ &\stackrel{①}{=} \frac{1}{b^2} \mathbb{E}_k[\sum_{i \in \mathcal{B}_k} \|X_i - \mathbb{E}_k[X_i]\|^2] \\ &\stackrel{②}{=} \frac{1}{b^2} \mathbb{E}_k[\sum_{i \in \mathcal{B}_k} \|G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)G_i w^k\|^2] - \frac{1}{b} [\mathbb{E}_k[X_i]]^2 \\ &= \frac{1}{nb} \sum_{i=1}^n \|G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)G_i w^k\|^2 - \frac{1}{b} [\mathbb{E}_k[X_i]]^2. \end{aligned}$$

Here, ① holds due to the i.i.d. property of \mathcal{B}_k , and ② holds since $\mathbb{E}_k[\|X_i - \mathbb{E}_k[X_i]\|^2] = \mathbb{E}_k[\|X_i\|^2] - (\mathbb{E}_k[X_i])^2$. This estimate implies the second line of (18) by taking the total expectation $\mathbb{E}[\cdot]$ both sides and the definition of Δ_k from (17).

Now, from (4) and (17), we can show that

$$\begin{aligned} \Delta_k &\stackrel{(17)}{:=} \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)G_i w^k\|^2] \\ &\stackrel{(4)}{=} \frac{(1-\mathbf{p})}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)G_i w^{k-1}\|^2] \\ &\quad + \frac{\mathbf{p}}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)G_i x^{k-1}\|^2] \\ &\stackrel{①}{\leq} \frac{(1+c)(1-\mathbf{p})}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^{k-1} - \gamma G_i x^{k-2} - (1-\gamma)G_i w^{k-1}\|^2] \\ &\quad + \frac{(1+c)(1-\mathbf{p})}{cnb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - [G_i x^{k-1} - \gamma G_i x^{k-2}]\|^2] \\ &\quad + \frac{\mathbf{p}}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - G_i x^{k-1}\|^2] \\ &\stackrel{②}{\leq} \frac{(1+c)(1-\mathbf{p})}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^{k-1} - \gamma G_i x^{k-2} - (1-\gamma)G_i w^{k-1}\|^2] \\ &\quad + \frac{2(1+c)(1-\mathbf{p})\gamma^2}{nbc} \sum_{i=1}^n \mathbb{E}[\|G_i x^{k-1} - G_i x^{k-2}\|^2] \\ &\quad + \frac{1}{nb} [\mathbf{p} + \frac{2(1+c)(1-\mathbf{p})}{c}] \sum_{i=1}^n \mathbb{E}[\|G_i x^k - G_i x^{k-1}\|^2] \\ &= (1+c)(1-\mathbf{p})\Delta_{k-1} + \frac{2(1+c)(1-\mathbf{p})\gamma^2}{nbc} \sum_{i=1}^n \mathbb{E}[\|G_i x^{k-1} - G_i x^{k-2}\|^2] \\ &\quad + \frac{1}{nb} [\mathbf{p} + \frac{2(1+c)(1-\mathbf{p})}{c}] \sum_{i=1}^n \mathbb{E}[\|G_i x^k - G_i x^{k-1}\|^2]. \end{aligned}$$

Here, in both inequalities ① and ②, we have used Young's inequality twice. If we choose $c := \frac{\mathbf{p}}{2(1-\mathbf{p})}$, then $(1+c)(1-\mathbf{p}) = 1 - \frac{\mathbf{p}}{2}$, $\frac{(1+c)(1-\mathbf{p})}{c} = (1-\mathbf{p})(1 + \frac{2(1-\mathbf{p})}{\mathbf{p}}) = \frac{(2-\mathbf{p})(1-\mathbf{p})}{\mathbf{p}} = \frac{2-3\mathbf{p}+\mathbf{p}^2}{\mathbf{p}}$, and $\frac{2(1+c)(1-\mathbf{p})\gamma^2}{nbc} + \mathbf{p} = \frac{4-6\mathbf{p}+3\mathbf{p}^2}{\mathbf{p}}$. Hence, we obtain

$$\begin{aligned} \Delta_k &\leq (1 - \frac{\mathbf{p}}{2})\Delta_{k-1} + \frac{(4-6\mathbf{p}+3\mathbf{p}^2)}{nb\mathbf{p}} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - G_i x^{k-1}\|^2] \\ &\quad + \frac{2\gamma^2(2-3\mathbf{p}+\mathbf{p}^2)}{nb\mathbf{p}} \sum_{i=1}^n \mathbb{E}[\|G_i x^{k-1} - G_i x^{k-2}\|^2]. \end{aligned}$$

This is exactly the last inequality of (18). \square

B.2. Proof of Lemma 2.3: SAGA estimator

Similarly, we also further expand Lemma 2.3 in detail as follows and then provide its full proof.

Lemma B.2. Let $S_\gamma^k := Gx^k - \gamma Gx^{k-1}$ be defined by (FRQ) and \tilde{S}_γ^k be generated by the SAGA estimator (SAGA), and $e^k := \tilde{S}_\gamma^k - S_\gamma^k$. We consider the following quantity:

$$\Delta_k := \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)\hat{G}_i^k\|^2]. \quad (19)$$

Then, we have

$$\begin{aligned}
 \mathbb{E}_k[\tilde{S}_\gamma^k] &= S_\gamma^k \equiv Gx^k - \gamma Gx^{k-1}, \\
 \mathbb{E}[\|\tilde{S}_\gamma^k - S_\gamma^k\|^2] &\leq \Delta_k - \frac{1}{b}\mathbb{E}[\|Gx^k - \gamma Gx^{k-1} - \frac{(1-\gamma)}{n} \sum_{i=1}^n \hat{G}_i^k\|^2] \leq \Delta_k, \\
 \Delta_k &\leq (1 - \frac{b}{2n})\Delta_{k-1} + \frac{[2(n-b)(2n+b)+b^2]}{n^2b^2} \sum_{i=1}^n \mathbb{E}[\|G_ix^k - G_ix^{k-1}\|^2] \\
 &\quad + \frac{2(n-b)(2n+b)\gamma^2}{n^2b^2} \sum_{i=1}^n \mathbb{E}[\|G_ix^{k-1} - G_ix^{k-2}\|^2].
 \end{aligned} \tag{20}$$

Consequently, the SAGA estimator \tilde{S}_γ^k constructed by (SAGA) satisfies Definition 2.1 with Δ_k in (19), $\rho := \frac{b}{2n} \in (0, 1]$, $C := \frac{[2(n-b)(2n+b)+b^2]}{nb^2}$, and $\hat{C} := \frac{2(n-b)(2n+b)\gamma^2}{nb^2}$.

Proof. It is well-known, see, e.g., (Defazio et al., 2014), that \tilde{S}_γ^k defined by (SAGA) is an unbiased estimator of S^k . Indeed, we have $\mathbb{E}_k[\hat{G}_{\mathcal{B}_k}^k] = \frac{1}{n} \sum_{i=1}^n \hat{G}_i^k$, $\mathbb{E}_k[G_{\mathcal{B}_k}x^k] = Gx^k$, and $\mathbb{E}_k[G_{\mathcal{B}_k}x^{k-1}] = Gx^{k-1}$. Using these relations and the definition of \tilde{S}_γ^k , we have

$$\begin{aligned}
 \mathbb{E}_k[\tilde{S}_\gamma^k] &= \mathbb{E}_k[\frac{(1-\gamma)}{n} \sum_{i=1}^n \hat{G}_i^k] - (1-\gamma)\mathbb{E}_k[\hat{G}_{\mathcal{B}_k}^k] + \mathbb{E}_k[G_{\mathcal{B}_k}x^k] - \gamma\mathbb{E}_k[G_{\mathcal{B}_k}x^{k-1}] \\
 &= \frac{(1-\gamma)}{n} \sum_{i=1}^n \hat{G}_i^k - \frac{(1-\gamma)}{n} \sum_{i=1}^n \hat{G}_i^k + Gx^k - \gamma Gx^{k-1} \\
 &= Gx^k - \gamma Gx^{k-1} \\
 &= S^k.
 \end{aligned}$$

Hence, \tilde{S}_γ^k is an unbiased estimator of S^k .

Next, let $X_i := G_ix^k - \gamma G_ix^{k-1} - (1-\gamma)\hat{G}_i^k$ for any $i \in [n]$. Then, we have $\mathbb{E}_k[X_i] = Gx^k - \gamma Gx^{k-1} - \frac{(1-\gamma)}{n} \sum_{i=1}^n \hat{G}_i^k$ for any $i \in [n]$. Therefore, we can derive

$$\begin{aligned}
 \mathbb{E}_k[\|\tilde{S}_\gamma^k - S_\gamma^k\|^2] &= \mathbb{E}_k[\|\frac{1}{b} \sum_{i \in \mathcal{B}_k} X_i - [Gx^k - \gamma Gx^{k-1} - \frac{(1-\gamma)}{n} \sum_{i=1}^n \hat{G}_i^k]\|^2] \\
 &= \mathbb{E}_k[\|\frac{1}{b} \sum_{i \in \mathcal{B}_k} X_i - \mathbb{E}_k[X_i]\|^2] \\
 &= \frac{1}{b^2} \mathbb{E}_k[\sum_{i \in \mathcal{B}_k} \|X_i - \mathbb{E}_k[X_i]\|^2] \\
 &= \frac{1}{b^2} \mathbb{E}_k[\sum_{i \in \mathcal{B}_k} \|G_ix^k - \gamma G_ix^{k-1} - (1-\gamma)\hat{G}_i^k\|^2] - \frac{1}{b} [\mathbb{E}_k[X_i]]^2 \\
 &= \frac{1}{nb} \sum_{i=1}^n \|G_ix^k - \gamma G_ix^{k-1} - (1-\gamma)\hat{G}_i^k\|^2 - \frac{1}{b} [\mathbb{E}_k[X_i]]^2.
 \end{aligned}$$

This implies the second line of (20) by taking the total expectation $\mathbb{E}[\cdot]$ both sides.

Now, from (5) and (19) and the rule (5), for any $c > 0$, by Young's inequality, we can show that

$$\begin{aligned}
 \Delta_k &\stackrel{(19)}{=} \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_ix^k - \gamma G_ix^{k-1} - (1-\gamma)\hat{G}_i^k\|^2] \\
 &\stackrel{(5)}{=} (1 - \frac{b}{n}) \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_ix^k - \gamma G_ix^{k-1} - (1-\gamma)\hat{G}_i^{k-1}\|^2] \\
 &\quad + \frac{b}{n} \cdot \frac{1}{nb} \sum_{i=1}^n \mathbb{E}[\|G_ix^k - \gamma G_ix^{k-1} - (1-\gamma)\hat{G}_i^{k-1}\|^2] \\
 &\leq \frac{(1+c)}{nb} (1 - \frac{b}{n}) \sum_{i=1}^n \mathbb{E}[\|G_ix^{k-1} - \gamma G_ix^{k-2} - (1-\gamma)\hat{G}_i^{k-1}\|^2] \\
 &\quad + \frac{(1+c)}{cnb} (1 - \frac{b}{n}) \sum_{i=1}^n \mathbb{E}[\|G_ix^k - \gamma G_ix^{k-1} - (G_ix^{k-1} - \gamma G_ix^{k-2})\|^2] \\
 &\quad + \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[\|G_ix^k - G_ix^{k-1}\|^2] \\
 &\leq (1+c)(1 - \frac{b}{n})\Delta_{k-1} + [\frac{1}{n^2} + (1 - \frac{b}{n}) \frac{2(1+c)}{cnb}] \sum_{i=1}^n \mathbb{E}[\|G_ix^k - G_ix^{k-1}\|^2] \\
 &\quad + \frac{2(1+c)\gamma^2}{cnb} (1 - \frac{b}{n}) \sum_{i=1}^n \mathbb{E}[\|G_ix^{k-1} - G_ix^{k-2}\|^2].
 \end{aligned}$$

If we choose $c := \frac{b}{2n} \in (0, 1)$, then $(1 - \frac{b}{n})(1+c) = 1 - \frac{b}{2n} - \frac{b^2}{2n^2} \leq 1 - \frac{b}{2n}$. Hence, we can further upper bound the last inequality as

$$\begin{aligned}
 \Delta_k &\leq (1 - \frac{b}{2n})\Delta_{k-1} + \frac{[2(n-b)(2n+b)+b^2]}{n^2b^2} \sum_{i=1}^n \mathbb{E}[\|G_ix^k - G_ix^{k-1}\|^2] \\
 &\quad + \frac{2(n-b)(2n+b)\gamma^2}{n^2b^2} \sum_{i=1}^n \mathbb{E}[\|G_ix^{k-1} - G_ix^{k-2}\|^2].
 \end{aligned}$$

This is exactly the last inequality of (20). \square

C. Convergence Analysis of VFR for (NE): Technical Proofs

To analyze our (VFR) scheme, we introduce the following two functions:

$$\begin{aligned}\mathcal{L}_k &:= \|x^k + \gamma\eta Gx^{k-1} - x^*\|^2 + \mu\|x^k - x^{k-1}\|^2, \\ \mathcal{E}_k &:= \mathcal{L}_k + \frac{\eta^2(1+\mu)(1-\rho)}{\rho}\Delta_{k-1} + \frac{L^2\eta^2\hat{C}(1+\mu)}{\rho}\|x^{k-1} - x^{k-2}\|^2,\end{aligned}\tag{21}$$

where μ is a given positive parameter, ρ , C , \hat{C} , and Δ_k are given in Definition 2.1, and $x^{-2} = x^{-1} = x^0$. Clearly, we have $\mathcal{L}_k \geq 0$ and $\mathcal{E}_k \geq 0$ for all $k \geq 0$ a.s.

One key step to analyze the convergence of (VFR) is to prove a descent property of \mathcal{E}_k defined by (21). The following lemma provides such a key estimate to prove the convergence of (VFR).

Lemma C.1. *Suppose that Assumptions 1.3 and 1.4 hold for (NE). Let $\{x^k\}$ be generated by (VFR) and \mathcal{E}_k be defined by (21) for any $\gamma \in [0, 1]$. Then, with $M := \frac{\gamma(1+\mu-\gamma)}{\mu} + \frac{(1+\mu)(C+\hat{C})}{\mu\rho}$, we have*

$$\begin{aligned}\mathbb{E}[\mathcal{E}_k] - \mathbb{E}[\mathcal{E}_{k+1}] &\geq \mu(1 - M \cdot L^2\eta^2)\mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad + \eta(1 - \gamma)[\eta(2\gamma - 1 - \mu) - 2\kappa]\mathbb{E}[\|Gx^k\|^2] \\ &\quad + \eta^2\gamma(1 - \gamma)(1 + \mu)\mathbb{E}[\|Gx^{k-1}\|^2].\end{aligned}\tag{22}$$

Proof. First, using $x^{k+1} := x^k - \eta\tilde{S}_\gamma^k$ from (VFR), we can expand

$$\begin{aligned}\|x^{k+1} + \gamma\eta Gx^k - x^*\|^2 &\stackrel{(\text{VFR})}{=} \|x^k - x^* + \gamma\eta Gx^k - \eta\tilde{S}_\gamma^k\|^2 \\ &= \|x^k - x^*\|^2 + 2\gamma\eta\langle Gx^k, x^k - x^* \rangle + \gamma^2\eta^2\|Gx^k\|^2 \\ &\quad - 2\eta\langle \tilde{S}_\gamma^k, x^k - x^* \rangle - 2\gamma\eta^2\langle Gx^k, \tilde{S}_\gamma^k \rangle + \eta^2\|\tilde{S}_\gamma^k\|^2.\end{aligned}$$

Second, it is obvious to show that

$$\|x^k + \gamma\eta Gx^{k-1} - x^*\|^2 = \|x^k - x^*\|^2 + 2\gamma\eta\langle Gx^{k-1}, x^k - x^* \rangle + \gamma^2\eta^2\|Gx^{k-1}\|^2.$$

Third, using again $x^{k+1} := x^k - \eta\tilde{S}_\gamma^k$ from (VFR), we can show that

$$\|x^{k+1} - x^k\|^2 = \eta^2\|\tilde{S}_\gamma^k\|^2.$$

Combining three expressions above, and using \mathcal{L}_k from (21), we can establish that

$$\begin{aligned}\mathcal{L}_k - \mathcal{L}_{k+1} &= \|x^k + \gamma\eta Gx^{k-1} - x^*\|^2 - \|x^{k+1} + \gamma\eta Gx^k - x^*\|^2 \\ &\quad + \mu\|x^k - x^{k-1}\|^2 - \mu\|x^{k+1} - x^k\|^2 \\ &= 2\gamma\eta\langle Gx^{k-1}, x^k - x^* \rangle - 2\gamma\eta\langle Gx^k, x^k - x^* \rangle + \gamma^2\eta^2\|Gx^{k-1}\|^2 \\ &\quad - \gamma^2\eta^2\|Gx^k\|^2 + 2\eta\langle \tilde{S}_\gamma^k, x^k - x^* \rangle + 2\gamma\eta^2\langle Gx^k, \tilde{S}_\gamma^k \rangle \\ &\quad + \mu\|x^k - x^{k-1}\|^2 - \eta^2(1 + \mu)\|\tilde{S}_\gamma^k\|^2.\end{aligned}\tag{23}$$

Next, since $\mathbb{E}_k[\tilde{S}_\gamma^k] = S_\gamma^k \equiv Gx^k - \gamma Gx^{k-1}$ as shown in the first line of (3) of Definition 2.1. Moreover, since \tilde{S}_γ^k is conditionally independent of $x^k - x^*$ and Gx^k w.r.t. the σ -field \mathcal{F}_k , we have

$$\begin{aligned}\mathbb{E}_k[\langle \tilde{S}_\gamma^k, x^k - x^* \rangle] &= \langle Gx^k, x^k - x^* \rangle - \gamma\langle Gx^{k-1}, x^k - x^* \rangle, \\ 2\mathbb{E}_k[\langle \tilde{S}_\gamma^k, Gx^k \rangle] &= 2\|Gx^k\|^2 - 2\gamma\langle Gx^{k-1}, Gx^k \rangle \\ &= (2 - \gamma)\|Gx^k\|^2 - \gamma\|Gx^{k-1}\|^2 + \gamma\|Gx^k - Gx^{k-1}\|^2.\end{aligned}$$

Taking the conditional expectation $\mathbb{E}_k[\cdot]$ both sides of (23) and using the last two expressions, we can show that

$$\begin{aligned}\mathcal{L}_k - \mathbb{E}_k[\mathcal{L}_{k+1}] &= 2\gamma\eta\langle Gx^{k-1}, x^k - x^* \rangle - 2\gamma\eta\langle Gx^k, x^k - x^* \rangle + \gamma^2\eta^2\|Gx^{k-1}\|^2 \\ &\quad - \gamma^2\eta^2\|Gx^k\|^2 + 2\eta\mathbb{E}_k[\langle \tilde{S}_\gamma^k, x^k - x^* \rangle] + 2\gamma\eta^2\mathbb{E}_k[\langle Gx^k, \tilde{S}_\gamma^k \rangle] \\ &\quad - \eta^2(1+\mu)\mathbb{E}_k[\|\tilde{S}_\gamma^k\|^2] + \mu\|x^k - x^{k-1}\|^2 \\ &= 2\eta(1-\gamma)\langle Gx^k, x^k - x^* \rangle + 2\gamma(1-\gamma)\eta^2\|Gx^k\|^2 \\ &\quad + \gamma^2\eta^2\|Gx^k - Gx^{k-1}\|^2 - \eta^2(1+\mu)\mathbb{E}_k[\|\tilde{S}_\gamma^k\|^2] + \mu\|x^k - x^{k-1}\|^2.\end{aligned}$$

Since \tilde{S}_γ^k is an unbiased estimator of S_γ^k , if we denote $e^k := \tilde{S}_\gamma^k - S_\gamma^k$, then we have $\mathbb{E}_k[e^k] = 0$. Hence, we can show that $\mathbb{E}_k[\|\tilde{S}_\gamma^k\|^2] = \mathbb{E}_k[\|S_\gamma^k + e^k\|^2] = \|S_\gamma^k\|^2 + 2\mathbb{E}_k[\langle e^k, S_\gamma^k \rangle] + \mathbb{E}_k[\|e^k\|^2] = \mathbb{E}_k[\|e^k\|^2] + \|S_\gamma^k\|^2$. Using this relation and $S_\gamma^k = Gx^k - \gamma Gx^{k-1}$, we can show that

$$\begin{aligned}\mathbb{E}_k[\|\tilde{S}_\gamma^k\|^2] &= \|S_\gamma^k\|^2 + \mathbb{E}_k[\|e^k\|^2] = \|Gx^k - \gamma Gx^{k-1}\|^2 + \mathbb{E}_k[\|e^k\|^2] \\ &= \|Gx^k\|^2 - 2\gamma\langle Gx^k, Gx^{k-1} \rangle + \gamma^2\|Gx^{k-1}\|^2 + \mathbb{E}_k[\|e^k\|^2] \\ &= (1-\gamma)\|Gx^k\|^2 - \gamma(1-\gamma)\|Gx^{k-1}\|^2 + \gamma\|Gx^k - Gx^{k-1}\|^2 + \mathbb{E}_k[\|e^k\|^2].\end{aligned}$$

Substituting this expression into the last estimate, we can show that

$$\begin{aligned}\mathcal{L}_k - \mathbb{E}_k[\mathcal{L}_{k+1}] &= 2\eta(1-\gamma)\langle Gx^k, x^k - x^* \rangle + \eta^2(1-\gamma)(2\gamma-1-\mu)\|Gx^k\|^2 \\ &\quad + \eta^2\gamma(1-\gamma)(1+\mu)\|Gx^{k-1}\|^2 - \gamma\eta^2(1+\mu-\gamma)\|Gx^k - Gx^{k-1}\|^2 \\ &\quad - \eta^2(1+\mu)\mathbb{E}_k[\|e^k\|^2] + \mu\|x^k - x^{k-1}\|^2.\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}[\mathcal{L}_k] - \mathbb{E}[\mathcal{L}_{k+1}] &= 2(1-\gamma)\eta\mathbb{E}[\langle Gx^k, x^k - x^* \rangle] + \eta^2\gamma(1-\gamma)(1+\mu)\mathbb{E}[\|Gx^{k-1}\|^2] \\ &\quad + \eta^2(1-\gamma)(2\gamma-1-\mu)\mathbb{E}[\|Gx^k\|^2] + \mu\mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad - \gamma\eta^2(1+\mu-\gamma)\mathbb{E}[\|Gx^k - Gx^{k-1}\|^2] - \eta^2(1+\mu)\mathbb{E}[\|e^k\|^2].\end{aligned}$$

By Young's inequality in ① and (2) of Assumption 1.3, we have

$$\begin{aligned}\|Gx^k - Gx^{k-1}\|^2 &= \|\frac{1}{n} \sum_{i=1}^n [G_i x^k - G_i x^{i-1}]\|^2 \stackrel{\textcircled{1}}{\leq} \frac{1}{n} \sum_{i=1}^n \|G_i x^k - G_i x^{i-1}\|^2 \\ &\stackrel{(2)}{\leq} L^2 \|x^k - x^{k-1}\|^2.\end{aligned}\tag{24}$$

Utilizing this inequality, $\langle Gx^k, x^k - x^* \rangle \geq -\kappa\|Gx^k\|^2$ from Assumption 1.4 with $T = 0$, and $\mathbb{E}[\|e^k\|^2] \leq \Delta_k$ from (3), we can bound the last expression as

$$\begin{aligned}\mathbb{E}[\mathcal{L}_k] - \mathbb{E}[\mathcal{L}_{k+1}] &\geq [\mu - L^2\eta^2\gamma(1+\mu-\gamma)]\mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad + (1+\mu)\gamma(1-\gamma)\eta^2\mathbb{E}[\|Gx^{k-1}\|^2] \\ &\quad + \eta(1-\gamma)[\eta(2\gamma-1-\mu) - 2\kappa]\mathbb{E}[\|Gx^k\|^2] - \eta^2(1+\mu)\Delta_k.\end{aligned}\tag{25}$$

By the third line of (3) in Definition 2.1 and again (2), we have

$$\Delta_k \leq (1-\rho)\Delta_{k-1} + CL^2\mathbb{E}[\|x^k - x^{k-1}\|^2] + \hat{C}L^2\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2].$$

Rearranging this inequality, we get

$$\begin{aligned}\Delta_k &\leq \left(\frac{1-\rho}{\rho}\right)(\Delta_{k-1} - \Delta_k) + \frac{\hat{C}L^2}{\rho}[\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2] - \mathbb{E}[\|x^k - x^{k-1}\|^2]] \\ &\quad + \frac{(C+\hat{C})L^2}{\rho}\mathbb{E}[\|x^k - x^{k-1}\|^2].\end{aligned}$$

Substituting this inequality into (25), we can show that

$$\begin{aligned}\mathbb{E}[\mathcal{L}_k] - \mathbb{E}[\mathcal{L}_{k+1}] &\geq \left[\mu - L^2\eta^2\gamma(1 + \mu - \gamma) - \frac{L^2\eta^2(1+\mu)(C+\hat{C})}{\rho} \right] \mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad + \eta(1 - \gamma)[\eta(2\gamma - 1 - \mu) - 2\kappa] \mathbb{E}[\|Gx^k\|^2] \\ &\quad + (1 + \mu)\gamma(1 - \gamma)\eta^2 \mathbb{E}[\|Gx^{k-1}\|^2] \\ &\quad - \frac{L^2\eta^2\hat{C}(1+\mu)}{\rho} [\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2] - \mathbb{E}[\|x^k - x^{k-1}\|^2]] \\ &\quad - \frac{\eta^2(1+\mu)(1-\rho)}{\rho} (\Delta_{k-1} - \Delta_k).\end{aligned}$$

Rearranging this inequality and using \mathcal{E}_k from (21), we obtain (22). \square

Now, we are ready to prove our first main result, Theorem 3.1 in the main text.

Proof of Theorem 3.1. Let us denote by $M := \frac{\gamma(1+\mu-\gamma)}{\mu} + \frac{(1+\mu)(C+\hat{C})}{\rho\mu}$. Then, to keep the right-hand side of (22) positive, we need to choose the parameters such that $L^2\eta^2 \leq \frac{1}{M}$ and $\eta \geq \frac{2\kappa}{2\gamma-1-\mu}$. These two conditions lead to $\frac{4L^2\kappa^2}{(2\gamma-1-\mu)^2} \leq L^2\eta^2 \leq \frac{1}{M}$.

Now, for a given $\gamma \in (\frac{1}{2}, 1)$, let us choose $\mu := \frac{3(2\gamma-1)}{4} > 0$. Then, the last condition holds if $L\kappa \leq \delta := \frac{2\gamma-1}{8\sqrt{M}}$ as stated in Theorem 3.1. In this case, we have $M = \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{1+6\gamma}{3(2\gamma-1)} \cdot \frac{C+\hat{C}}{\rho}$ as stated in (6). Hence, we can choose $\frac{8\kappa}{2\gamma-1} \leq \eta \leq \frac{1}{L\sqrt{M}}$ as claimed in Theorem 3.1.

Next, utilizing $\mu + 1 = \frac{1+6\gamma}{2} \geq 1$ and $\mu = \frac{3(2\gamma-1)}{4}$, (22) reduces to

$$\mathbb{E}[\mathcal{E}_k] - \mathbb{E}[\mathcal{E}_{k+1}] \geq \frac{3(2\gamma-1)}{4} (1 - M \cdot L^2\eta^2) \mathbb{E}[\|x^k - x^{k-1}\|^2] + \gamma(1 - \gamma)\eta^2 \mathbb{E}[\|Gx^{k-1}\|^2].$$

Averaging this inequality from $k := 0$ to $k := K$, we obtain

$$\begin{cases} \frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^{k-1}\|^2] &\leq \frac{\mathbb{E}[\mathcal{E}_0]}{\gamma(1-\gamma)\eta^2(K+1)}, \\ \frac{(1-ML^2\eta^2)}{K+1} \sum_{k=0}^K \mathbb{E}[\|x^k - x^{k-1}\|^2] &\leq \frac{4\mathbb{E}[\mathcal{E}_0]}{3(2\gamma-1)(K+1)}.\end{cases}$$

Finally, since $x^{-1} = x^{-2} = x^0$, and \tilde{S}_γ^0 is chosen as $\tilde{S}_\gamma^0 := (1-\gamma)Gx^0$, we have $\Delta_{-1} = \Delta_0 = 0$. Using this fact, $Gx^* = 0$, the Lipschitz continuity of G , $\rho \in [0, 1]$, and $\gamma < 1$, we can show that

$$\begin{aligned}\mathbb{E}[\mathcal{E}_0] &= \mathbb{E}[\|x^0 + \eta\gamma G(x^0) - x^*\|^2] + \frac{\eta^2(1+\mu)(1-\rho)}{\rho} \Delta_0 \\ &\leq 2\mathbb{E}[\|x^0 - x^*\|^2] + 2\eta^2\gamma^2 \mathbb{E}[\|Gx^0 - Gx^*\|^2] + \frac{(1+6\gamma)\eta^2}{4\rho} \Delta_0 \\ &\leq 2(1 + L^2\eta^2\gamma^2) \mathbb{E}[\|x^0 - x^*\|^2] + \frac{(1+6\gamma)\eta^2}{4\rho} \Delta_0 \\ &\leq 2(1 + L^2\eta^2) \|x^0 - x^*\|^2.\end{aligned}$$

Substituting this upper bound into the above estimates, we get the second bound of (7). For the first bound, we replace $k-1$ by k , and K by $K+1$, using $\|Gx^0\|^2 \leq 2\|Gx^0\|^2$, and then multiplying both sides of the result by $\frac{K+2}{K+1}$ to obtain the first line of (7). \square

Next, we restate Corollary 3.2 for the case $\gamma \in (\frac{1}{2}, 1)$ instead of $\gamma = \frac{3}{4}$ as in the main text. Then, we derive the proof of Corollary 3.2 from this result by fixing $\gamma = \frac{3}{4}$.

Corollary C.2. Suppose that Assumptions 1.1, 1.3, and 1.4 hold for (NE) with $\kappa \geq 0$ as in Theorem 3.1. Let $\{x^k\}$ be generated by (VFR) using the SVRG estimator (L-SVRG), $\gamma \in (\frac{1}{2}, 1)$, and

$$\eta := \frac{1}{L\sqrt{M}}, \quad \text{where} \quad \Lambda := \frac{4(1+\gamma^2)(2-3\rho)+2(3+2\gamma^2)\rho^2}{b\rho^2} \quad \text{and} \quad M := \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{1+6\gamma}{3(2\gamma-1)} \cdot \Lambda. \quad (26)$$

Then, we have $\eta \geq \frac{\sigma\sqrt{b}\mathbf{p}}{L}$ for $\sigma := \frac{\sqrt{3(2\gamma-1)}}{\sqrt{8+49\gamma+13\gamma^2+48\gamma^3}}$, and the following bound holds:

$$\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \frac{2(1+L^2\eta^2)R_0^2}{\gamma(1-\gamma)\eta^2(K+1)}, \quad \text{where } R_0 := \|x^0 - x^*\|. \quad (27)$$

For a given tolerance $\epsilon > 0$, if we choose $\mathbf{p} := n^{-1/3}$ and $b := \lfloor n^{2/3} \rfloor$, then (VFR) requires $\mathcal{T}_{G_i} := n + \lfloor \frac{4\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \rfloor$ evaluations of G_i to achieve $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \epsilon^2$, where $\Gamma := \frac{2(5\gamma^2+7\gamma-3)(8+49\gamma+13\gamma^2+48\gamma^3)}{3\gamma^2(2\gamma-1)(1-\gamma)(1+5\gamma)}$.

Proof. For the SVRG estimator (L-SVRG), by Lemma 2.2, we have $\rho := \frac{\mathbf{p}}{2} \in (0, 1]$, $C := \frac{4-6\mathbf{p}+3\mathbf{p}^2}{b\mathbf{p}}$, and $\hat{C} := \frac{2\gamma^2(2-3\mathbf{p}+\mathbf{p}^2)}{b\mathbf{p}}$. Therefore, we can compute $\Lambda := \frac{C+\hat{C}}{\rho} = \frac{4(1+\gamma^2)(2-3\mathbf{p})+2(3+2\gamma^2)\mathbf{p}^2}{b\mathbf{p}^2} \leq \frac{8(1+\gamma^2)}{b\mathbf{p}^2}$, and M in Theorem 3.1 as $M := \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{1+6\gamma}{3(2\gamma-1)} \cdot \Lambda \leq \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{8(1+6\gamma)(1+\gamma^2)}{3(2\gamma-1)b\mathbf{p}^2}$ as stated in (26). The estimate (27) is exactly the first line of (7).

Now, suppose that $b\mathbf{p}^2 \leq 1$. Then, by (26), we have $M \leq \frac{8+49\gamma+13\gamma^2+48\gamma^3}{3(2\gamma-1)b\mathbf{p}^2}$. Therefore, if we choose $\eta := \frac{1}{L\sqrt{M}}$, then η satisfies the conditions of Theorem 3.1, provided that $L\rho \leq \delta$. Moreover, we have $\eta \geq \frac{\sqrt{3(2\gamma-1)}\sqrt{b}\mathbf{p}}{\sqrt{8+49\gamma+13\gamma^2+48\gamma^3}} = \frac{\sigma\sqrt{b}\mathbf{p}}{L}$, where $\sigma := \frac{\sqrt{3(2\gamma-1)}}{\sqrt{8+49\gamma+13\gamma^2+48\gamma^3}}$.

From (27), to guarantee $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \epsilon^2$, we need to impose $\frac{2(1+L^2\eta^2)R_0^2}{\gamma(1-\gamma)\eta^2(K+1)} \leq \epsilon^2$, where $R_0 := \|x^0 - x^*\|$. However, since $1 + L^2\eta^2 = 1 + \frac{1}{M} \leq \frac{5\gamma^2+7\gamma-3}{\gamma(1+5\gamma)}$ and $\eta \geq \frac{\sigma\sqrt{b}\mathbf{p}}{L}$, the last condition holds if we choose $K := \left\lceil \Gamma \cdot \frac{L^2 R_0^2}{b\mathbf{p}^2 \epsilon^2} \right\rceil$, where $\Gamma := \frac{2(5\gamma^2+7\gamma-3)}{\sigma^2\gamma^2(1-\gamma)(1+5\gamma)} = \frac{2(5\gamma^2+7\gamma-3)(8+49\gamma+13\gamma^2+48\gamma^3)}{3\gamma^2(2\gamma-1)(1-\gamma)(1+5\gamma)}$.

Finally, note that, at each iteration k , (VFR) requires 3 mini-batches of size b , and occasionally compute the full batch Gw^k , leading to the cost of $n\mathbf{p} + 3b$. The total complexity is

$$\mathcal{T}_c := K(n\mathbf{p} + 3b) = \frac{\Gamma L^2 R_0^2 (n\mathbf{p} + 3b)}{b\mathbf{p}^2 \epsilon^2} = \frac{\Gamma L^2 R_0^2}{\epsilon^2} \left(\frac{n}{b\mathbf{p}} + \frac{3}{\mathbf{p}^2} \right).$$

If we choose $b := \lfloor n^{2/3} \rfloor$ and $\mathbf{p} := n^{-1/3}$, then $b\mathbf{p}^2 = 1$ and $\mathcal{T}_c = \frac{4\Gamma n^{2/3} L^2 R_0^2}{\epsilon^2}$. For the SVRG estimator (L-SVRG), one needs to compute Gw^0 , which requires n evaluations of G_i . Hence, the total complexity of the algorithm is $\mathcal{T}_{G_i} := n + \left\lceil \frac{4\Gamma n^{2/3} L^2 R_0^2}{\epsilon^2} \right\rceil$ as stated. \square

Proof of Corollary 3.2. Since we fix $\gamma := \frac{3}{4}$, we can easily compute $\sigma := \frac{\sqrt{3(2\gamma-1)}}{\sqrt{8+49\gamma+13\gamma^2+48\gamma^3}} \approx 0.144025 \geq 0.1440$ and $\Gamma := \frac{2(5\gamma^2+7\gamma-3)(8+49\gamma+13\gamma^2+48\gamma^3)}{3\gamma^2(2\gamma-1)(1-\gamma)(1+5\gamma)} \approx 730.736842 \leq 731$. Therefore, we obtain $\eta \geq \frac{0.1440\sqrt{b}\mathbf{p}}{L}$ and $\mathcal{T}_{G_i} := n + \left\lceil \frac{4\Gamma n^{2/3} L^2 R_0^2}{\epsilon^2} \right\rceil$, where $\Gamma := 731$. Moreover, (27) reduces to $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \frac{32(1+0.1440^2)L^2 R_0^2}{3 \cdot 0.1440^2 b\mathbf{p}^2 (K+1)} \leq \frac{526 \cdot L^2 R_0^2}{b\mathbf{p}^2 (K+1)}$. \square

Finally, we also restate Corollary 3.3 for the case $\gamma \in (\frac{1}{2}, 1)$ and then derive the oracle complexity of Corollary 3.3 from this result by fixing $\gamma := \frac{3}{4}$.

Corollary C.3. Suppose that Assumptions 1.1, 1.3, and 1.4 hold for (NE) with $\kappa \geq 0$ as in Theorem 3.1. Let $\{x^k\}$ be generated by (VFR) using the SAGA estimator (SAGA), $\gamma \in (\frac{1}{2}, 1)$, and

$$\eta := \frac{1}{L\sqrt{M}}, \quad \text{where } \Lambda := \frac{2}{b} + \frac{4(1+\gamma^2)(n-b)(2n+b)}{b^3} \quad \text{and} \quad M := \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{1+6\gamma}{3(2\gamma-1)} \cdot \Lambda. \quad (28)$$

Then, we have $\eta \geq \frac{\sigma b^{3/2}}{nL}$ for $\sigma := \frac{\sqrt{3(2\gamma-1)}}{\sqrt{10+61\gamma+13\gamma^2+48\gamma^3}}$, and the following bound holds:

$$\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \frac{2(1+L^2\eta^2)R_0^2}{\gamma(1-\gamma)\eta^2(K+1)}, \quad \text{where } R_0 := \|x^0 - x^*\|. \quad (29)$$

Moreover, for a given tolerance $\epsilon > 0$, if we choose $b := \lfloor n^{2/3} \rfloor$, then (VFR) requires $\mathcal{T}_{G_i} := n + \left\lceil \frac{3\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \right\rceil$ evaluations of G_i to achieve $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \epsilon^2$, where $\Gamma := \frac{2(7\gamma+5\gamma^2-3)(10+61\gamma+13\gamma^2+48\gamma^3)}{3\gamma^2(1-\gamma)(2\gamma-1)(1+5\gamma)}$.

Proof. Since we use the SAGA estimator (SAGA), we have $\rho := \frac{b}{2n} \in (0, 1]$, $C := \frac{[2(n-b)(2n+b)+b^2]}{nb^2}$, and $\hat{C} := \frac{2(n-b)(2n+b)\gamma^2}{nb^2}$. In this case, since $b \geq 1$, we can easily show that $\Lambda := \frac{C+\hat{C}}{\rho} = \frac{2}{b} + \frac{4(1+\gamma^2)(n-b)(2n+b)}{b^3} \leq 2 + \frac{8(1+\gamma^2)n^2}{b^3}$. Hence, M in Theorem 3.1 reduces to

$$M := \frac{\gamma(1+5\gamma)}{3(2\gamma-1)} + \frac{1+6\gamma}{3(2\gamma-1)} \cdot \Lambda \leq \frac{2+13\gamma+5\gamma^2}{3(2\gamma-1)} + \frac{8(1+\gamma^2)(1+6\gamma)n^2}{3(2\gamma-1)b^3}.$$

Suppose that $1 \leq b \leq n^{2/3}$. Then, one can prove that $M \leq \left[\frac{2+13\gamma+5\gamma^2}{3(2\gamma-1)} + \frac{8(1+\gamma^2)(1+6\gamma)}{3(2\gamma-1)} \right] \frac{n^2}{b^3} = \frac{(10+61\gamma+13\gamma^2+48\gamma^3)n^2}{3(2\gamma-1)b^3} = \frac{n^2}{\sigma^2 b^3}$, where $\sigma := \frac{\sqrt{3(2\gamma-1)}}{\sqrt{10+61\gamma+13\gamma^2+48\gamma^3}}$. Hence, if we choose $\eta := \frac{1}{L\sqrt{M}}$, then we get $\eta \geq \frac{\sigma b^{3/2}}{nL}$ as stated. Moreover, we obtain (27) from the first line of (7) as before.

Now, for $\eta := \frac{1}{L\sqrt{M}} \geq \frac{\sigma b^{3/2}}{nL}$, from (27), to guarantee $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E} [\|Gx^k\|^2] \leq \epsilon^2$, we need to impose $\frac{2(1+L^2\eta^2)R_0^2}{\gamma(1-\gamma)\eta^2(K+1)} \leq \epsilon^2$, where $R_0 := \|x^0 - x^*\|$. Since $1 + L^2\eta^2 = 1 + \frac{1}{M} \leq \frac{7\gamma+5\gamma^2-3}{\gamma(1+5\gamma)}$ and $\eta \geq \frac{\sigma b^{3/2}}{nL}$, the last condition holds if we choose $K := \left\lceil \Gamma \cdot \frac{L^2 R_0^2 n^2}{b^3 \epsilon^2} \right\rceil$, where $\Gamma := \frac{2(7\gamma+5\gamma^2-3)}{\sigma^2 \gamma^2 (1-\gamma)(1+5\gamma)} = \frac{2(7\gamma+5\gamma^2-3)(10+61\gamma+13\gamma^2+48\gamma^3)}{3\gamma^2(1-\gamma)(2\gamma-1)(1+5\gamma)}$.

Finally, at each iteration k , (VFR) requires 3 mini-batches of size b , leading to the cost of $3b$ per iteration. Hence, the total complexity is

$$\mathcal{T}_c := 3bK = \left\lceil \frac{3\Gamma L^2 R_0^2 n^2}{b^2 \epsilon^2} \right\rceil.$$

If we choose $b := \lfloor n^{2/3} \rfloor$, then $\mathcal{T}_c = \left\lceil \frac{3\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \right\rceil$. For the SAGA estimator (SAGA), one needs to compute Gw^0 , which requires n evaluations of G_i . Hence, the total complexity of the algorithm is $\mathcal{T}_{G_i} := n + \left\lceil \frac{3\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \right\rceil$. \square

Proof of Corollary 3.3. Since we choose $\gamma := \frac{3}{4}$, we have $\sigma := \frac{\sqrt{3(2\gamma-1)}}{\sqrt{10+61\gamma+13\gamma^2+48\gamma^3}} = 0.14948 \geq 0.1494$ and $\Gamma := \frac{2(7\gamma+5\gamma^2-3)(10+61\gamma+13\gamma^2+48\gamma^3)}{3\gamma^2(1-\gamma)(2\gamma-1)(1+5\gamma)} = 2815.8 \leq 2816$. Applying the results of Corollary C.3, we obtain our conclusions in Corollary 3.3. Moreover, (29) reduces to $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E} [\|Gx^k\|^2] \leq \frac{32(1+0.494^2)L^2 R_0^2}{3 \cdot 0.1494^2 b \mathbf{p}^2 (K+1)} \leq \frac{489 \cdot L^2 R_0^2}{b \mathbf{p}^2 (K+1)}$ as stated. \square

D. Convergence Analysis of VFRBS for (NI): Technical Proofs

One key step to analyze the convergence of (VFRBS) is to construct an appropriate potential function. For this purpose, we introduce the following function:

$$\mathcal{L}_k := \|x^k + \gamma\eta(Gx^{k-1} + v^k) - x^*\|^2 + \mu\|x^k - x^{k-1} + \gamma\eta(Gx^{k-1} + v^k)\|^2, \quad (30)$$

where $\mu > 0$ is a given parameter and $v^k \in Tx^k$ is given. This function is then combined with \mathcal{E}_k from (21) to establish the convergence of (VFRBS).

Let us first state and prove Lemma D.1, which provides a key estimate for our convergence analysis of (VFRBS) in Theorem 4.1.

Lemma D.1. Suppose that Assumption 1.3 holds for (NI). Let $\{x^k\}$ be generated by (VFRBS), \mathcal{L}_k be defined by (30), and \mathcal{E}_k be defined by (21). Then, we have

$$\begin{aligned} \mathcal{L}_k - \mathbb{E}_k[\mathcal{L}_{k+1}] &\geq 2(1-\gamma)\eta\langle Gx^k + v^k, x^k - x^* \rangle + (1+\mu)(1-\gamma)(2\gamma-1)\eta^2\|Gx^k + v^k\|^2 \\ &\quad + \gamma[1-\gamma-\mu(3\gamma-1)]\eta^2\|Gx^{k-1} + v^k\|^2 - (1+\mu)\eta^2\mathbb{E}_k[\|e^k\|^2] \\ &\quad + \frac{1}{2}[\mu - 2(1+\mu)\gamma(1-\gamma)L^2\eta^2]\|x^k - x^{k-1}\|^2. \end{aligned} \quad (31)$$

If, additionally, Assumption 1.4 holds for (NI), then we have

$$\begin{aligned} \mathbb{E}[\mathcal{E}_k] - \mathbb{E}[\mathcal{E}_{k+1}] &\geq \frac{1}{2}\left[\mu - 2(1+\mu)\gamma(1-\gamma)L^2\eta^2 - \frac{2L^2\eta^2(1+\mu)(C+\hat{C})}{\rho}\right]\mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad + \gamma[1-\gamma-\mu(3\gamma-1)]\eta^2\mathbb{E}[\|Gx^{k-1} + v^k\|^2] \\ &\quad + (1-\gamma)\eta[(1+\mu)(2\gamma-1)\eta - 2\kappa]\mathbb{E}[\|Gx^k + v^k\|^2]. \end{aligned} \quad (32)$$

Proof. Let us introduce two notations $w^k := Gx^k + v^k$ and $\hat{w}^k := Gx^{k-1} + v^k$, where $v^k \in Tx^k$. We also recall $S_\gamma^k := Gx^k - \gamma Gx^{k-1}$ and $e^k := \tilde{S}_\gamma^k - S_\gamma^k$ from (FRQ). Then, it is obvious that $\tilde{S}_\gamma^k = S_\gamma^k + e^k = Gx^k - \gamma Gx^{k-1} + e^k$.

Now, using $\tilde{S}_\gamma^k = Gx^k - \gamma Gx^{k-1} + e^k$, it follows from (VFRBS) that

$$\begin{aligned} x^{k+1} &= x^k - \eta \tilde{S}_\gamma^k - \gamma \eta v^{k+1} - (2\gamma - 1)\eta v^k \\ &= x^k - \gamma \eta (Gx^k + v^{k+1}) - (1 - \gamma)\eta (Gx^k + v^k) + \eta \gamma (Gx^{k-1} + v^k) - \eta e^k \\ &= x^k - \gamma \eta \hat{w}^{k+1} - (1 - \gamma)\eta w^k + \gamma \eta \hat{w}^k - \eta e^k. \end{aligned} \quad (33)$$

Then, using (33) and $\hat{w}^{k+1} = Gx^k + v^{k+1}$, we can show that

$$\begin{aligned} \mathcal{T}_{[1]} &:= \|x^{k+1} + \gamma \eta (Gx^k + v^{k+1}) - x^*\|^2 = \|x^{k+1} - x^* + \gamma \eta \hat{w}^{k+1}\|^2 \\ &\stackrel{(33)}{=} \|x^k - \gamma \eta \hat{w}^{k+1} - (1 - \gamma)\eta w^k + \gamma \eta \hat{w}^k - \eta e^k - x^* + \gamma \eta \hat{w}^{k+1}\|^2 \\ &= \|x^k - x^*\|^2 - 2(1 - \gamma)\eta \langle w^k, x^k - x^* \rangle + 2\gamma \eta \langle \hat{w}^k, x^k - x^* \rangle + \eta^2 \|e^k\|^2 \\ &\quad + (1 - \gamma)^2 \eta^2 \|w^k\|^2 - 2\gamma(1 - \gamma)\eta^2 \langle w^k, \hat{w}^k \rangle + \gamma^2 \eta^2 \|\hat{w}^k\|^2 \\ &\quad - 2\eta \langle e^k, x^k - x^* \rangle + 2(1 - \gamma)\eta^2 \langle e^k, w^k \rangle - 2\gamma \eta^2 \langle e^k, \hat{w}^k \rangle. \end{aligned}$$

Alternatively, using $\hat{w}^k = Gx^{k-1} + v^k$, we also have

$$\begin{aligned} \mathcal{T}_{[2]} &:= \|x^k + \gamma \eta (Gx^{k-1} + v^k) - x^*\|^2 = \|x^k - x^* + \gamma \eta \hat{w}^k\|^2 \\ &= \|x^k - x^*\|^2 + 2\gamma \eta \langle \hat{w}^k, x^k - x^* \rangle + \gamma^2 \eta^2 \|\hat{w}^k\|^2. \end{aligned}$$

Subtracting $\mathcal{T}_{[1]}$ from $\mathcal{T}_{[2]}$, we can show that

$$\begin{aligned} \mathcal{T}_{[3]} &:= \|x^k + \gamma \eta (Gx^{k-1} + v^k) - x^*\|^2 - \|x^{k+1} + \gamma \eta (Gx^k + v^{k+1}) - x^*\|^2 \\ &= 2(1 - \gamma)\eta \langle w^k, x^k - x^* \rangle - (1 - \gamma)^2 \eta^2 \|w^k\|^2 + 2\gamma(1 - \gamma)\eta^2 \langle w^k, \hat{w}^k \rangle \\ &\quad + 2\eta \langle e^k, x^k - x^* \rangle - 2(1 - \gamma)\eta^2 \langle e^k, w^k \rangle + 2\gamma \eta^2 \langle e^k, \hat{w}^k \rangle - \eta^2 \|e^k\|^2 \\ &= 2(1 - \gamma)\eta \langle w^k, x^k - x^* \rangle + (1 - \gamma)(2\gamma - 1)\eta^2 \|w^k\|^2 \\ &\quad + \gamma(1 - \gamma)\eta^2 \|\hat{w}^k\|^2 - \gamma(1 - \gamma)\eta^2 \|w^k - \hat{w}^k\|^2 \\ &\quad + 2\eta \langle e^k, x^k - x^* \rangle - 2(1 - \gamma)\eta^2 \langle e^k, w^k \rangle + 2\gamma \eta^2 \langle e^k, \hat{w}^k \rangle - \eta^2 \|e^k\|^2. \end{aligned} \quad (34)$$

Next, using again $\hat{w}^{k+1} = Gx^k + v^{k+1}$ and (33), we have

$$\begin{aligned} \mathcal{T}_{[4]} &:= \|x^{k+1} - x^k + \gamma \eta (Gx^k + v^{k+1})\|^2 = \|x^{k+1} - x^k + \gamma \eta \hat{w}^{k+1}\|^2 \\ &\stackrel{(33)}{=} \eta^2 \|(1 - \gamma)w^k - \gamma \hat{w}^k + e^k\|^2 \\ &= (1 - \gamma)^2 \eta^2 \|w^k\|^2 - 2\gamma(1 - \gamma)\eta^2 \langle w^k, \hat{w}^k \rangle + \gamma^2 \eta^2 \|\hat{w}^k\|^2 \\ &\quad + \eta^2 \|e^k\|^2 + 2(1 - \gamma)\eta^2 \langle e^k, w^k \rangle - 2\gamma \eta^2 \langle e^k, \hat{w}^k \rangle \\ &= -(1 - \gamma)(2\gamma - 1)\eta^2 \|w^k\|^2 + \gamma(2\gamma - 1)\eta^2 \|\hat{w}^k\|^2 + \gamma(1 - \gamma)\eta^2 \|w^k - \hat{w}^k\|^2 \\ &\quad + \eta^2 \|e^k\|^2 + 2(1 - \gamma)\eta^2 \langle e^k, w^k \rangle - 2\gamma \eta^2 \langle e^k, \hat{w}^k \rangle. \end{aligned}$$

Moreover, by the Cauchy-Schwarz inequality in ① and Young's inequality in ②, we can prove that

$$\begin{aligned} \|x^k - x^{k-1} + \gamma \eta \hat{w}^k\|^2 &= \|x^k - x^{k-1}\|^2 + 2\gamma \eta \langle \hat{w}^k, x^k - x^{k-1} \rangle + \gamma^2 \eta^2 \|\hat{w}^k\|^2 \\ &\stackrel{①}{\geq} \|x^k - x^{k-1}\|^2 - 2\gamma \eta \|\hat{w}^k\| \|x^k - x^{k-1}\| + \gamma^2 \eta^2 \|\hat{w}^k\|^2 \\ &\stackrel{②}{\geq} \frac{1}{2} \|x^k - x^{k-1}\|^2 - \gamma^2 \eta^2 \|\hat{w}^k\|^2. \end{aligned}$$

Combining the last two expressions, we can show that

$$\begin{aligned}
 \mathcal{T}_{[5]} &:= \|x^k - x^{k-1} + \gamma\eta(Gx^{k-1} + v^k)\|^2 - \|x^{k+1} - x^k + \gamma\eta(Gx^k + v^{k+1})\|^2 \\
 &= \|x^k - x^{k-1} + \gamma\eta\hat{w}^k\|^2 - \|x^{k+1} - x^k + \gamma\eta\hat{w}^{k+1}\|^2 \\
 &\geq \frac{1}{2}\|x^k - x^{k-1}\|^2 + (1-\gamma)(2\gamma-1)\eta^2\|w^k\|^2 - \gamma(3\gamma-1)\eta^2\|\hat{w}^k\|^2 \\
 &\quad - \gamma(1-\gamma)\eta^2\|w^k - \hat{w}^k\|^2 - \eta^2\|e^k\|^2 - 2(1-\gamma)\eta^2\langle e^k, w^k \rangle + 2\gamma\eta^2\langle e^k, \hat{w}^k \rangle.
 \end{aligned}$$

Multiplying $\mathcal{T}_{[5]}$ by $\mu > 0$, and adding the result to (34), and using \mathcal{L}_k from (30), we have

$$\begin{aligned}
 \mathcal{L}_k - \mathcal{L}_{k+1} &= \|x^k + \gamma\eta(Gx^{k-1} + v^k) - x^*\|^2 - \|x^{k+1} + \gamma\eta(Gx^k + v^{k+1}) - x^*\|^2 \\
 &\quad + \mu\|x^k - x^{k-1} + \gamma\eta(Gx^{k-1} + v^k)\|^2 - \mu\|x^{k+1} - x^k + \gamma\eta(Gx^k + v^{k+1})\|^2 \\
 &\geq 2(1-\gamma)\eta\langle w^k, x^k - x^* \rangle + \frac{\mu}{2}\|x^k - x^{k-1}\|^2 + (1+\mu)(1-\gamma)(2\gamma-1)\eta^2\|w^k\|^2 \\
 &\quad + \gamma[(1-\gamma) - \mu(3\gamma-1)]\eta^2\|\hat{w}^k\|^2 - (1+\mu)\gamma(1-\gamma)\eta^2\|w^k - \hat{w}^k\|^2 \\
 &\quad + 2\eta\langle e^k, x^k - x^* \rangle - 2(1+\mu)(1-\gamma)\eta^2\langle e^k, w^k \rangle \\
 &\quad + 2(1+\mu)\gamma\eta^2\langle e^k, \hat{w}^k \rangle - (1+\mu)\eta^2\|e^k\|^2.
 \end{aligned}$$

Taking the conditional expectation $\mathbb{E}_k[\cdot]$ both sides of this expression, and noting that

$$\begin{aligned}
 \mathbb{E}_k[\langle e^k, x^k - x^* \rangle] &= \langle \mathbb{E}_k[e^k], x^k - x^* \rangle = 0, \\
 \mathbb{E}_k[\langle e^k, w^k \rangle] &= \langle \mathbb{E}_k[e^k], w^k \rangle = 0, \\
 \mathbb{E}_k[\langle e^k, \hat{w}^k \rangle] &= \langle \mathbb{E}_k[e^k], \hat{w}^k \rangle = 0,
 \end{aligned}$$

we obtain

$$\begin{aligned}
 \mathcal{L}_k - \mathbb{E}_k[\mathcal{L}_{k+1}] &\geq 2(1-\gamma)\eta\langle w^k, x^k - x^* \rangle + \frac{\mu}{2}\|x^k - x^{k-1}\|^2 + (1+\mu)(1-\gamma)(2\gamma-1)\eta^2\|w^k\|^2 \\
 &\quad + \gamma[(1-\gamma) - \mu(3\gamma-1)]\eta^2\|\hat{w}^k\|^2 - (1+\mu)\gamma(1-\gamma)\eta^2\|w^k - \hat{w}^k\|^2 \\
 &\quad - (1+\mu)\eta^2\mathbb{E}_k[\|e^k\|^2].
 \end{aligned}$$

Finally, by the L -Lipschitz continuity of G from (2) of Assumption 1.3, we have $\|w^k - \hat{w}^k\|^2 = \|Gx^k - Gx^{k-1}\|^2 \leq L^2\|x^k - x^{k-1}\|^2$ as shown in (24). Using this inequality into the last estimate, we can show that

$$\begin{aligned}
 \mathcal{L}_k - \mathbb{E}_k[\mathcal{L}_{k+1}] &\geq 2(1-\gamma)\eta\langle w^k, x^k - x^* \rangle + (1+\mu)(1-\gamma)(2\gamma-1)\eta^2\|w^k\|^2 \\
 &\quad + \gamma[1-\gamma-\mu(3\gamma-1)]\eta^2\|\hat{w}^k\|^2 - (1+\mu)\eta^2\mathbb{E}_k[\|e^k\|^2] \\
 &\quad + \frac{1}{2}[\mu - 2(1+\mu)\gamma(1-\gamma)L^2\eta^2]\|x^k - x^{k-1}\|^2,
 \end{aligned}$$

which proves (31) by recalling $w^k := Gx^k + v^k$ and $\hat{w}^k := Gx^{k-1} + v^k$.

Taking the full expectation of (31) and using $\langle Gx^k + v^k, x^k - x^* \rangle \geq -\kappa\|Gx^k + v^k\|^2$ from Assumption 1.4 and $\mathbb{E}_k[\|e^k\|^2] \leq \Delta_k$ from (3), we can bound it as

$$\begin{aligned}
 \mathbb{E}[\mathcal{L}_k] - \mathbb{E}[\mathcal{L}_{k+1}] &\geq \frac{1}{2}[\mu - 2(1+\mu)\gamma(1-\gamma)L^2\eta^2]\mathbb{E}[\|x^k - x^{k-1}\|^2] - (1+\mu)\eta^2\Delta_k \\
 &\quad + \gamma[1-\gamma-\mu(3\gamma-1)]\eta^2\mathbb{E}[\|Gx^{k-1} + v^k\|^2] \\
 &\quad + (1-\gamma)\eta[(1+\mu)(2\gamma-1)\eta - 2\kappa]\mathbb{E}[\|Gx^k + v^k\|^2].
 \end{aligned} \tag{35}$$

By the third line of (3) in Definition 2.1 and utilizing again (2), we have

$$\Delta_k \leq (1-\rho)\Delta_{k-1} + CL^2\mathbb{E}[\|x^k - x^{k-1}\|^2] + \hat{C}L^2\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2].$$

Rearranging this inequality, we get

$$\begin{aligned}\Delta_k &\leq \left(\frac{1-\rho}{\rho}\right)(\Delta_{k-1} - \Delta_k) + \frac{\hat{C}L^2}{\rho} [\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2] - \mathbb{E}[\|x^k - x^{k-1}\|^2]] \\ &\quad + \frac{(C+\hat{C})L^2}{\rho} \mathbb{E}[\|x^k - x^{k-1}\|^2].\end{aligned}$$

Substituting this inequality into (35), we can show that

$$\begin{aligned}\mathbb{E}[\mathcal{L}_k] - \mathbb{E}[\mathcal{L}_{k+1}] &\geq \frac{1}{2} \left[\mu - 2(1+\mu)\gamma(1-\gamma)L^2\eta^2 - \frac{2L^2\eta^2(1+\mu)(C+\hat{C})}{\rho} \right] \mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad + \gamma[1-\gamma-\mu(3\gamma-1)]\eta^2\mathbb{E}[\|Gx^{k-1} + v^k\|^2] \\ &\quad + (1-\gamma)\eta[(1+\mu)(2\gamma-1)\eta-2\kappa]\mathbb{E}[\|Gx^k + v^k\|^2] \\ &\quad - \frac{L^2\eta^2\hat{C}(1+\mu)}{\rho} \left[\mathbb{E}[\|x^{k-1} - x^{k-2}\|^2] - \mathbb{E}[\|x^k - x^{k-1}\|^2] \right] \\ &\quad - \frac{\eta^2(1+\mu)(1-\rho)}{\rho} (\Delta_{k-1} - \Delta_k).\end{aligned}$$

Rearranging this inequality and using \mathcal{E}_k from (21), we obtain (32). \square

Now, we are ready to prove our second main result, Theorem 4.1 in the main text.

Proof of Theorem 4.1. Since we fix $\gamma \in (\frac{1}{2}, 1)$ and $\mu := \frac{1-\gamma}{3\gamma-1}$, we have $\mu > 0$ and $1 + \mu = \frac{2\gamma}{3\gamma-1}$. Let us denote by $M := 4\gamma^2 + \frac{4\gamma}{1-\gamma} \cdot \frac{C+\hat{C}}{\rho}$ as in Theorem 4.1. Then, (32) reduces to

$$\begin{aligned}\mathbb{E}[\mathcal{E}_k] - \mathbb{E}[\mathcal{E}_{k+1}] &\geq \frac{(1-\gamma)(1-M\cdot L^2\eta^2)}{2(3\gamma-1)} \mathbb{E}[\|x^k - x^{k-1}\|^2] \\ &\quad + 2(1-\gamma)\eta \left[\frac{\gamma(2\gamma-1)\eta}{3\gamma-1} - \kappa \right] \mathbb{E}[\|Gx^k + v^k\|^2].\end{aligned}\tag{36}$$

Let us choose $\eta > 0$ such that $\frac{\gamma(2\gamma-1)\eta}{3\gamma-1} - \kappa > 0$ and $1 - M \cdot L^2\eta^2 \geq 0$. These two conditions lead to $\frac{(3\gamma-1)\kappa}{\gamma(2\gamma-1)} < \eta \leq \frac{1}{L\sqrt{M}}$ as stated in Theorem 4.1. However, this condition holds if $L^2\kappa^2 < \frac{\gamma^2(2\gamma-1)^2}{M(3\gamma-1)^2}$. This condition is equivalent to $L\kappa \leq \delta$ as our condition in Theorem 4.1, where $\delta := \frac{\gamma(2\gamma-1)}{(3\gamma-1)\sqrt{M}}$.

Averaging (36) from $k = 0$ to K and noting that $\mathbb{E}[\mathcal{E}_k] \geq 0$ for all $k \geq 0$, we get

$$\begin{aligned}\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k + v^k\|^2] &\leq \frac{(3\gamma-1) \cdot \mathbb{E}[\mathcal{E}_0]}{2(1-\gamma)[\gamma(2\gamma-1)\eta - (3\gamma-1)\kappa]\eta(K+1)}, \\ \frac{(1-M\cdot L^2\eta^2)}{K+1} \sum_{k=0}^K \mathbb{E}[\|x^k - x^{k-1}\|^2] &\leq \frac{2(3\gamma-1) \cdot \mathbb{E}[\mathcal{E}_0]}{(1-\gamma)(K+1)}.\end{aligned}$$

Finally, since $x^{-1} = x^{-2} = x^0$, we have $\Delta_{-1} = \Delta_0$. However, since $\tilde{S}_\gamma^0 = (1-\gamma)Gx^0 = S_\gamma^0$, we get $\Delta_0 = \|\tilde{S}_\gamma^0 - S_\gamma^0\|^2 = 0$. Using these relations, $\rho \in [0, 1]$ and $\gamma < 1$, we can show that

$$\begin{aligned}\mathbb{E}[\mathcal{E}_0] &= \mathbb{E}[\|x^0 + \gamma\eta(Gx^0 + v^0) - x^*\|^2] + \frac{\eta^2(1+\mu)(1-\rho)}{\rho} \Delta_0 \\ &\leq 2\mathbb{E}[\|x^0 - x^*\|^2] + 2\gamma^2\eta^2\mathbb{E}[\|Gx^0 + v^0\|^2] + \frac{2\gamma\eta^2}{(3\gamma-1)\rho} \Delta_0 \\ &= 2\mathbb{E}[\|x^0 - x^*\|^2] + 2\gamma^2\eta^2\mathbb{E}[\|Gx^0 + v^0\|^2].\end{aligned}$$

Substituting this upper bound into the above two estimates, we get two lines of (12). \square

Finally, we prove Corollaries 4.2 and 4.3 in the main text. Unlike Corollaries 3.2 and 3.3 where we fix $\gamma := \frac{3}{4}$, here we state these corollaries for any value of $\gamma \in (\frac{1}{2}, 1)$.

Proof of Corollary 4.2. For the SVRG estimator (L-SVRG), we have $\rho := \frac{\mathbf{p}}{2} \in (0, 1]$, $C := \frac{4-6\mathbf{p}+3\mathbf{p}^2}{b\mathbf{p}}$, $\hat{C} := \frac{2\gamma^2(2-3\mathbf{p}+\mathbf{p}^2)}{b\mathbf{p}}$, and $\Delta_0 = 0$ due to (17) and $x^0 = x^{-1} = w^0$. In this case, we have $\Lambda := \frac{C+\hat{C}}{\rho} = \frac{4(1+\gamma^2)(2-3\mathbf{p})+2(3+2\gamma^2)\mathbf{p}^2}{b\mathbf{p}^2} \leq \frac{8(1+\gamma^2)}{b\mathbf{p}^2}$, and thus M in Theorem 3.1 reduces to $M := 4\gamma^2 + \frac{4\gamma}{1-\gamma}\Lambda \leq 4\gamma^2 + \frac{32(1+\gamma^2)}{b\mathbf{p}^2}$.

Suppose that $b\mathbf{p}^2 \leq 1$. Since $\Lambda \leq \frac{8(1+\gamma^2)}{b\mathbf{p}^2}$ and $M = 4\gamma^2 + \frac{4\gamma}{1-\gamma}\Lambda \leq 4\gamma^2 + \frac{32\gamma(1+\gamma^2)}{(1-\gamma)b\mathbf{p}^2} \leq \frac{4\gamma(8+\gamma+7\gamma^2)}{(1-\gamma)b\mathbf{p}^2}$. If we choose $\eta := \frac{1}{L\sqrt{M}}$, then we have $\eta \geq \frac{\sqrt{1-\gamma}\sqrt{b\mathbf{p}}}{2L\sqrt{\gamma(8+\gamma+7\gamma^2)}} = \frac{\sigma\sqrt{b\mathbf{p}}}{L}$ with $\sigma := \frac{\sqrt{1-\gamma}}{2\sqrt{8+\gamma+7\gamma^2}}$, then it satisfies $\frac{(3\gamma-1)\kappa}{\gamma(2\gamma-1)} < \eta \leq \frac{1}{L\sqrt{M}}$ in Theorem 4.1, provided that $L\kappa \leq \delta$. Note that using $\eta \geq \frac{\sigma\sqrt{b\mathbf{p}}}{L}$ in (12) of Theorem 4.1 we obtain the bound (13).

Now, from the first line of (12), to guarantee $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E} [\|Gx^k + v^k\|^2] \leq \epsilon^2$, we need to impose $\frac{\Theta \hat{R}_0^2}{\eta^2(K+1)} \leq \epsilon^2$, where $\hat{R}_0^2 := \|x^0 - x^*\|^2 + \gamma^2\eta^2\|Gx^0 + v^0\|^2$. Since $\eta \geq \frac{\sigma\sqrt{b\mathbf{p}}}{L}$, the last condition holds if we choose $K := \left\lceil \Gamma \cdot \frac{L^2 \hat{R}_0^2}{b\mathbf{p}^2 \epsilon^2} \right\rceil$, where $\Gamma := \frac{\Theta}{\sigma^2}$.

Finally, at each iteration k , (VFRBS) requires 3 mini-batches of size b , and occasionally compute the full Gw^k , leading to the cost of $n\mathbf{p} + 3b$ per iteration. Thus the total complexity is

$$\mathcal{T}_c := K(n\mathbf{p} + 3b) = \frac{\Gamma L^2 \hat{R}_0^2 (n\mathbf{p} + 3b)}{b\mathbf{p}^2 \epsilon^2} = \frac{\Gamma L^2 \hat{R}_0^2}{\epsilon^2} \left(\frac{n}{b\mathbf{p}} + \frac{3}{\mathbf{p}^2} \right).$$

If we choose $b := \lfloor n^{2/3} \rfloor$ and $\mathbf{p} := n^{-1/3}$, then $b\mathbf{p}^2 = 1$ and $\mathcal{T}_c = \frac{4\Gamma n^{2/3} L^2 \hat{R}_0^2}{\epsilon^2}$. For the SVRG estimator (L-SVRG), one needs to compute Gw^0 , which requires n evaluations of G_i . Hence, the total evaluations of G_i is $\mathcal{T}_{G_i} = n + \left\lfloor \frac{4\Gamma n^{2/3} L^2 \hat{R}_0^2}{\epsilon^2} \right\rfloor$. Moreover, at each iteration, we need one evaluation of $J_{\gamma\eta T}$. Therefore, the total evaluations of $J_{\gamma\eta T}$ is $\mathcal{T}_T := K = \left\lfloor \Gamma \cdot \frac{L^2 \hat{R}_0^2}{b\mathbf{p}^2 \epsilon^2} \right\rfloor = \left\lfloor \Gamma \cdot \frac{L^2 \hat{R}_0^2}{\epsilon^2} \right\rfloor$. \square

Proof of Corollary 4.3. Since we use the SAGA estimator (SAGA), we have $\rho := \frac{b}{2n} \in (0, 1]$, $C := \frac{[2(n-b)(2n+b)+b^2]}{nb^2}$, and $\hat{C} := \frac{2(n-b)(2n+b)\gamma^2}{nb^2}$. In this case, since $b \geq 1$, we get $\Lambda := \frac{C+\hat{C}}{\rho} = \frac{2}{b} + \frac{4(1+\gamma^2)(n-b)(2n+b)}{b^3} \leq 2 + \frac{8(1+\gamma^2)n^2}{b^3}$. Hence, M in Theorem 3.1 reduces to

$$M := 4\gamma^2 + \frac{4\gamma}{1-\gamma} \cdot \Lambda \leq \frac{4\gamma(2+\gamma-\gamma^2)}{1-\gamma} + \frac{32\gamma(1+\gamma^2)n^2}{(1-\gamma)b^3}$$

Suppose that $1 \leq b \leq n^{2/3}$. Then, we can show that $M \leq \left[\frac{4\gamma(2+\gamma-\gamma^2)}{1-\gamma} + \frac{32\gamma(1+\gamma^2)}{1-\gamma} \right] \frac{n^2}{b^3} = \frac{4\gamma(10+\gamma+7\gamma^2)}{(1-\gamma)b^3} = \frac{n^2}{\sigma^2 b^3}$, where $\sigma := \frac{\sqrt{1-\gamma}}{2\sqrt{\gamma(10+\gamma+7\gamma^2)}}$. Hence, if we choose $\eta := \frac{1}{L\sqrt{M}}$, then we get $\eta \geq \frac{\sigma b^{3/2}}{nL}$. Note that using $\eta \geq \frac{\sigma b^{3/2}}{nL}$ in (12) of Theorem 4.1 we obtain the bound (14).

For $\eta := \frac{1}{L\sqrt{M}} \geq \frac{\sigma b^{3/2}}{nL}$, from the first line of (12), to guarantee $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E} [\|Gx^k + v^k\|^2] \leq \epsilon^2$, we need to impose $\frac{\Theta \hat{R}_0^2}{\eta^2(K+1)} \leq \epsilon^2$, where $\hat{R}_0^2 := \|x^0 - x^*\|^2 + \gamma^2\eta^2\|Gx^0 + v^0\|^2$. Since $\eta \geq \frac{\sigma b^{3/2}}{nL}$, the last condition holds if we choose $K := \left\lceil \Gamma \cdot \frac{L^2 \hat{R}_0^2 n^2}{b^3 \epsilon^2} \right\rceil$, where $\Gamma := \frac{\Theta}{\sigma^2}$.

Finally, at each iteration k , (VFRBS) requires 3 mini-batches of size b , leading to the cost of $3b$ per iteration. Thus the total complexity is

$$\mathcal{T}_c := 3bK = \left\lceil \frac{3\Gamma L^2 \hat{R}_0^2 n^2}{b^2 \epsilon^2} \right\rceil.$$

If we choose $b := \lfloor n^{2/3} \rfloor$, then $\mathcal{T}_c = \left\lceil \frac{3\Gamma L^2 \hat{R}_0^2 n^{2/3}}{\epsilon^2} \right\rceil$. For the SAGA estimator (SAGA), one needs to compute Gw^0 , which requires n evaluations of G_i . We conclude that (VFRBS) requires $\mathcal{T}_{G_i} := n + \left\lfloor \frac{3\Gamma L^2 \hat{R}_0^2 n^{2/3}}{\epsilon^2} \right\rfloor$ evaluations of G_i . Moreover, since each iteration, it requires one evaluation of $J_{\gamma\eta T}$, we need $\mathcal{T}_T := K = \left\lfloor \Gamma \cdot \frac{L^2 \hat{R}_0^2}{\epsilon^2} \right\rfloor$ evaluations of $J_{\gamma\eta T}$. \square

Remark D.2. For the SVRG estimator, if we choose $\gamma = \frac{3}{4}$, then we have $\sigma := 0.0702$. Hence, we have $\eta \geq \frac{0.0702\sqrt{b\mathbf{p}}}{L}$. However, if we choose $\gamma := 0.55$, then $\eta \geq \frac{0.1027\sqrt{b\mathbf{p}}}{L}$. If we choose $b = \lfloor n^{2/3} \rfloor$ and $\mathbf{p} = n^{-1/3}$, then the latter lower bound becomes $\eta \geq \frac{0.1027}{L}$.

For the SAGA estimator, if we choose $\gamma = \frac{3}{4}$, then we have $\sigma := 0.0753$. Hence, we get $\eta \geq \frac{0.0753b^{3/2}}{nL}$. However, if we set $\gamma := 0.55$, then $\eta \geq \frac{0.1271b^{3/2}}{nL}$. If we choose $b = \lfloor n^{2/3} \rfloor$, then the latter lower bound becomes $\eta \geq \frac{0.1271}{L}$.

Note that these lower bounds of η can be further improved by refining the related parameters in Lemma D.1, and carefully choosing μ in the proof of Theorem 4.1.

E. Details of Experiments and Additional Experiments

Due to space limit, we do not provide the details of experiments in Section 5. In this Supp. Doc., we provide the details of our implementation and experiments. We also add more examples to illustrate our algorithms and compare them with existing methods. All algorithms are implemented in Python, and all the experiments are run on a MacBookPro. 2.8GHz Quad-Core Intel Core I7, 16Gb Memory.

E.1. Synthetic WGAN Example

We modify the synthetic example in (Daskalakis et al., 2018) built up on WGAN from (Arjovsky et al., 2017) as our first example. Suppose that the generator is a simple additive model $G_\theta(z) = \theta + z$ with the noise input z generated from a normal distribution $\mathcal{N}(0, \mathbb{I})$, and the discriminator is also a linear function $D_\beta(w) = \langle K\beta, w \rangle$ for a given matrix K , where $\theta \in \mathbb{R}^{p_1}$ and $\beta \in \mathbb{R}^{p_2}$, and $K \in \mathbb{R}^{p_1 \times p_2}$ is a given matrix. The goal of the generator is to find a true distribution $\theta = \theta^*$, leading to the following loss:

$$\mathcal{L}(\theta, \beta) := \mathbb{E}_{u \sim \mathcal{N}(\theta^*, \mathbb{I})} [\langle K\beta, u \rangle] - \mathbb{E}_{z \sim \mathcal{N}(0, \mathbb{I})} [\langle K\beta, \theta + z \rangle].$$

Suppose that we have n samples for both w and z leading to the following bilinear minimax problem:

$$\inf_{\theta \in \mathbb{R}^{p_1}} \sup_{\beta \in \mathbb{R}^{p_2}} \left\{ \mathcal{L}(\theta, \beta) := f(\theta) + \frac{1}{n} \sum_{i=1}^n [\langle K\beta, w_i - z_i - \theta \rangle] - g(\beta) \right\}. \quad (37)$$

Here, we add two convex functions $f(\theta)$ and $g(\beta)$ to possibly handle constraints or regularizers associated with θ and β , respectively.

If we define $x := [\theta, \beta] \in \mathbb{R}^{p_1+p_2}$, $Gx = [\nabla_\theta \mathcal{L}(\theta, \beta), -\nabla_\beta \mathcal{L}(\theta, \beta)] := -[K\beta, \frac{1}{n} \sum_{i=1}^n K^\top (w_i - z_i - \theta)]$, and $T := [\partial f(\theta), \partial g(\beta)]$, then the optimality condition of this minimax problem becomes $0 \in Gx + Tx$, which is a special case of (NI) with Gx being linear. The model (37) is different from the one in (Daskalakis et al., 2018) at two points:

- It involves a linear operator K , making it more general than (Daskalakis et al., 2018).
- It has two additional terms f and g , making it broader to also cover constraints or non-smooth regularizers.

In our experiments below, we consider two cases:

- **Case 1 (Unconstrained setting).** We assume that $\theta \in \mathbb{R}^{p_1}$ and $\beta \in \mathbb{R}^{p_2}$.
- **Case 2 (Constrained setting).** Assume that θ and β stays in an ℓ_∞ -ball of radius $r > 0$, leading to $f(\theta) := \delta_{[-r, r]^{p_1}}(\theta)$ and $g(\beta) := \delta_{[-r, r]^{p_2}}(\beta)$, the indicator of the ℓ_∞ -balls.

E.1.1. THE UNCONSTRAINED CASE

(a) **Algorithms.** We implement three variants of (VFR) to solve (37).

- The first variant is using a double-loop SVRG strategy (called VFR-SVRG), where the full operator Gw^s at a snapshot point w^s is computed at the beginning of each epoch s . Then, we perform $\lfloor n/b \rfloor$ iterations k to update x^k using (VFR), where b is the mini-batch size. Finally, we set the next snapshot point $w^{s+1} := x^{k+1}$ after finishing the inner loop.
- The second variant is called a loopless one, LVFR-SVRG, where we implement exactly the same scheme (VFR) as in this paper and using the Loopless-SVRG estimator.
- The third variant is VFR-SAGA, where we use the SAGA estimator in (VFR).

We also compare our methods with the deterministic optimistic gradient (OG) in (Daskalakis et al., 2018), the variance-reduced FRBS (VFRBS) in (Alacaoglu et al., 2023), and the variance-reduced extragradient (VEG) in (Alacaoglu & Malitsky, 2022).

(b) **Input data.** For (NE), we generate a vector θ^* from the standard normal distribution as our true mean in \mathbb{R}^{p_1} . Then, we generate i.i.d. samples w_i and z_i from normal distribution $\mathcal{N}(\theta^*, \mathbb{I})$ and $\mathcal{N}(0, \mathbb{I})$, respectively for $i = 1, 2, \dots, n$ in \mathbb{R}^{p_1} and \mathbb{R}^{p_2} , respectively. We perform two experiments: **Experiment 1** with $n = 5000$ and $p_1 = p_2 = 100$, and **Experiment 2** with $n = 10000$ and $p_1 = p_2 = 200$. For each experiment, we run 10 times up to 100 epochs, corresponding to 10 problem instances, using the same setting, but different input data (w_i, z_i) , and then compute the mean of the relative operator norm $\|Gx^k\|/\|Gx^0\|$. This mean is then plotted.

(c) **Parameters.** For the optimistic gradient algorithm (OG), we choose its learning rate $\eta := \frac{1}{L}$, where L is the Lipschitz constant of G , though its theoretical learning rate is much smaller. For our methods in (VFR), if $n = 5000$, and we choose $b := \lfloor 0.5n^{2/3} \rfloor = 146$, and the probability $\mathbf{p} := \frac{2}{n^{1/3}} = 0.1170$, then $\eta := \frac{1}{L\sqrt{M}} = \frac{0.1905}{L}$. However, due to the under

estimation of M , we instead use a larger learning rate $\eta := \frac{1}{2L}$ for all three variants, and choose a mini-batch of size $b := \lfloor 0.5n^{2/3} \rfloor$, and a probability $\mathbf{p} := \frac{1}{n^{1/3}}$ for the loopless SVRG variant.

For the forward-reflected-backward splitting method with variance reduction (VFRBS) in (Alacaoglu et al., 2023), we choose its learning rate $\eta := \frac{0.95(1-\sqrt{1-\mathbf{p}})}{2L}$ as suggested by (Alacaoglu et al., 2023). However, we still choose the probability $\mathbf{p} = \frac{1}{n^{1/3}}$ and the mini-batch size $b = \lfloor 0.5n^{2/3} \rfloor$ as our methods. These values are much larger the ones suggested in (Alacaoglu et al., 2023), typically $\mathbf{p} = \mathcal{O}(1/n)$.

For the variance reduction extragradient method (VEG) in (Alacaoglu & Malitsky, 2022), we choose its learning rate $\eta := \frac{0.95\sqrt{1-\alpha}}{L}$ for $\alpha := 1 - \mathbf{p}$ from the paper. However, again, we also choose $\mathbf{p} := \frac{1}{n^{1/3}}$ and $b = \lfloor 0.5n^{2/3} \rfloor$ in this method, which is the same as ours, though their theoretical results suggest smaller values of \mathbf{p} (e.g., $\mathbf{p} = \frac{1}{n}$). Note that if $n = 5000$, then the batch size $b := 150$ and the probability $\mathbf{p} := 0.062$, but if $n = 10000$, then $b = 239$ and $\mathbf{p} = 0.0479$.

(d) **Experiments for $K = \mathbb{I}$.** We perform two experiments: **Experiment 1** with $(n, p) = (5000, 200)$ and **Experiment 2** with $(n, p) = (10000, 400)$ as discussed above. We run each experiment with 10 problem instances and compute the mean of the relative residual norm $\|Gx^k\|/\|Gx^0\|$. The results of this test are plotted in Figure 4.

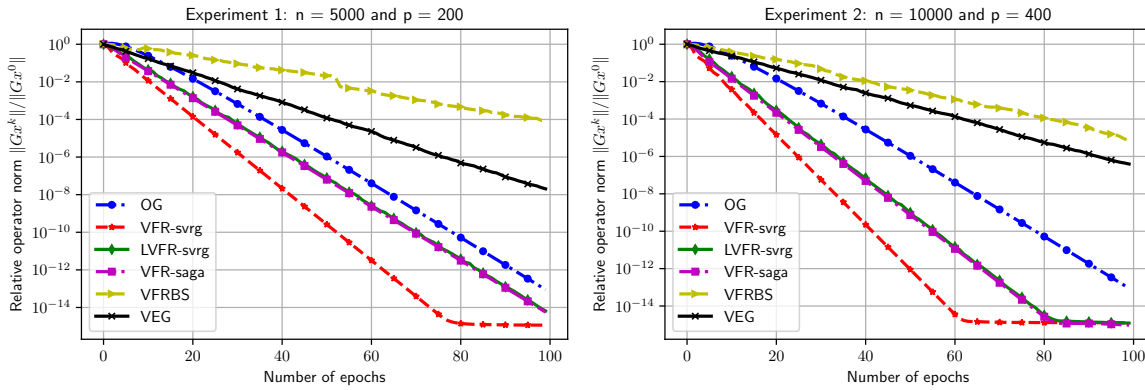


Figure 4: Performance of 6 algorithms to solve (37) on 2 experiments when $K = \mathbb{I}$.

For these particular experiments, our methods highly outperform OG, VFRBS, and VEG. It shows that VFR-svrg is the best overall, while LVFR-saga and VFR-svrg have a similar performance in both experiments. Both the competitors: VFRBS and VEG do not perform well in this test and they are much slower than ours and also OG. This is perhaps due to a small learning rate of VFRBS although we choose the same mini-batch size b and the same probability \mathbf{p} as ours.

(e) **Experiments for $K \neq \mathbb{I}$.** Now, we test these 6 algorithms for the case $K \neq \mathbb{I}$ in our extended model (37), where K is generated randomly from the standard normal distribution. Then, we normalize K as $K/\|K\|$ to get a unit Lipschitz constant $L = 1$.

Again, we use the same configuration as in Figure 4 and also run our experiments on 10 problems and report the mean results. We perform two experiments: **Experiment 1** with $n = 5000$ and $p_1 = p_2 = p = 100$, and **Experiment 2** with $n = 10000$ and $p_1 = p_2 = p = 200$. The results are reported in Figure 5.

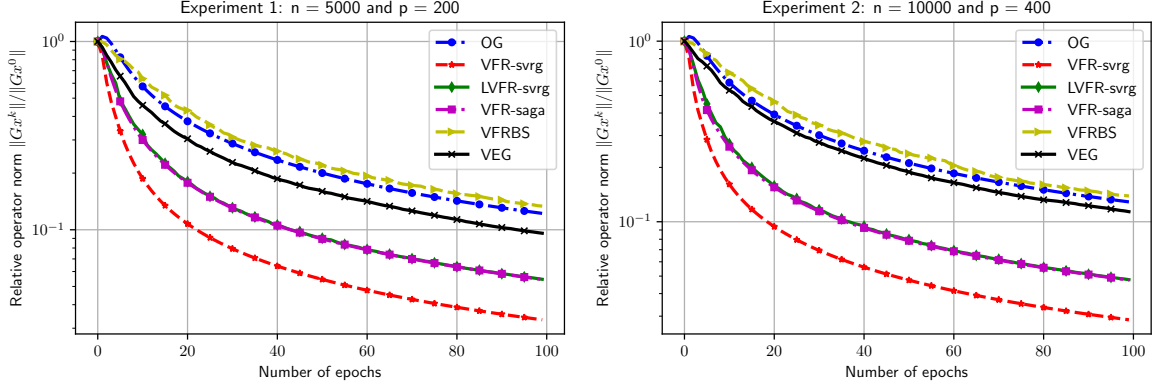
We still observe that our algorithms work well and outperform their competitors. However, after 100 epochs, these methods can only reach a 10^{-2} accuracy level for an approximate solution.

E.1.2. THE UNCONSTRAINED CASE – VARYING b AND \mathbf{p}

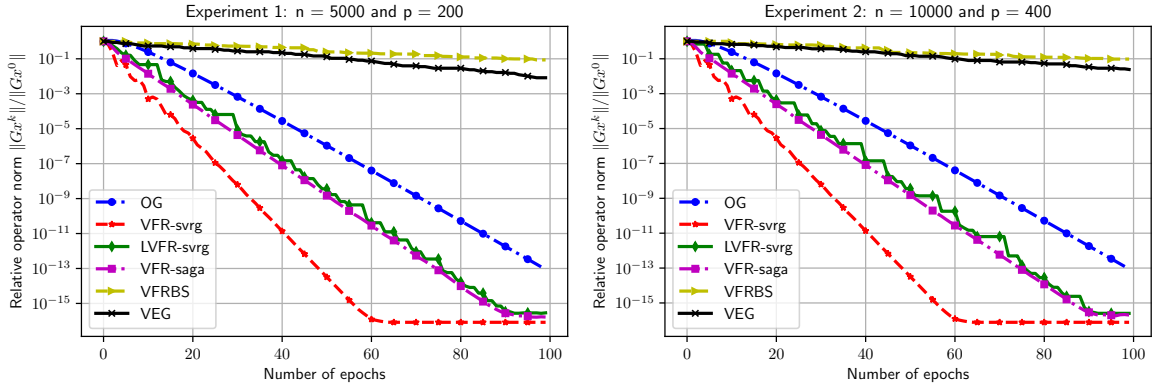
We can certainly tune the parameters to make our competitors (VFRBS) and (VEG) work better. However, such parameter configurations are far from satisfying the conditions of their theoretical results. For example, if we set $\mathbf{p} = \frac{20}{\sqrt{n}}$, then both VFRBS and VEG work better. In particular, if $n = 5000$, then we get $\mathbf{p} = \frac{20}{\sqrt{n}} = 0.28$, which is several times larger than its suggested value $\mathbf{p} = \frac{1}{n} = 2 \times 10^{-4}$.

Let us further experiment other choices of parameters (i.e. the mini-batch size b and the probability \mathbf{p} of flipping a coin) to observe the performance of these algorithms.

(a) **Larger b .** Figure 6 reveals the performance of these algorithms when we increase the mini-batch size b to a larger value

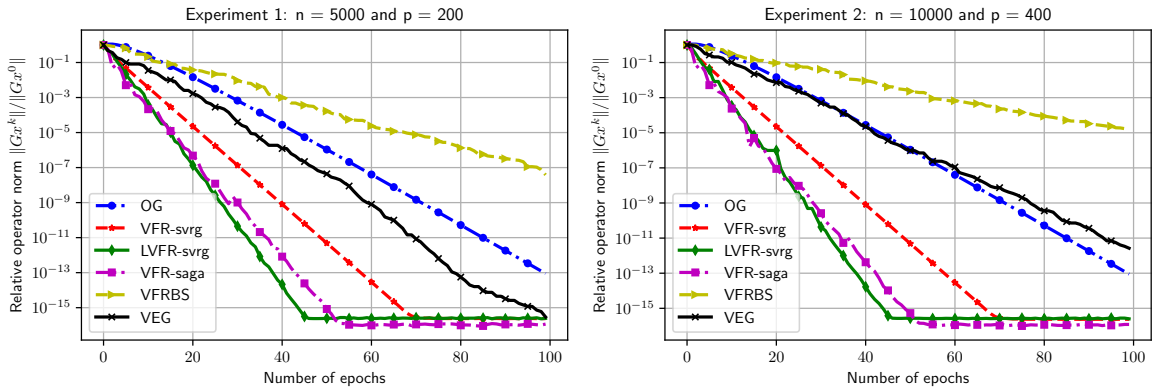

 Figure 5: Performance of 6 algorithms to solve (37) on 2 experiments when $K \neq \mathbb{I}$.

$b = \lfloor 0.1n \rfloor$, while keeping the probability $\mathbf{p} = \frac{1}{n^{1/3}}$ unchanged.


 Figure 6: Performance of 6 algorithms for a large $b = \lfloor 0.1n \rfloor$ and a unchanged $\mathbf{p} = \frac{1}{n^{1/3}}$.

Note that for $n = 5000$, we have $b = 500$ and $\mathbf{p} = 0.058$, and for $n = 10000$, we have $b = 1000$ and $\mathbf{p} = 0.046$. With these large mini-batches, our algorithms still outperform other methods, while VFRBS and VEG are significantly slowed down. The double-loop variant of (VFR) with SVRG performs best, while LVFR-svrg and VFR-saga have a similar performance.

(b) **Medium b and larger \mathbf{p} .** Next, we set b to a medium size of $b = \lfloor 0.05n \rfloor$ (corresponding to $b = 250$ for $n = 5000$ and $b = 500$ for $n = 10000$) and increase $\mathbf{p} = \frac{1}{n^{1/4}}$ (corresponding to $\mathbf{p} = 0.119$ for $n = 5000$ and $\mathbf{p} = 0.1$ for $n = 10000$). Then, the results are shown in Figure 7.


 Figure 7: Performance of 6 algorithms for a medium $b = \lfloor 0.05n \rfloor$ and larger $\mathbf{p} = \frac{1}{n^{1/4}}$.

Then, we observe that LVFR-svrg and VFR-saga superiorly outperform the others. The performance of the double-loop VFR-svrg is still similar to the previous tests since it is not affected by \mathbf{p} . In addition, VEG is now comparable with OG, but VFRBS remains the slowest one.

(c) **Large b and small \mathbf{p} .** To see the effect of \mathbf{p} on our competitors: VFRBS and VEG, as suggested by their theory, we decrease \mathbf{p} to $\mathbf{p} = \frac{1}{n^{1/2}}$ (corresponding to $\mathbf{p} = 0.014$ for $n = 5000$ and $\mathbf{p} = 0.01$ for $n = 10000$) and still set $b = \lfloor 0.1n \rfloor$, and the results are plotted in Figure 8.

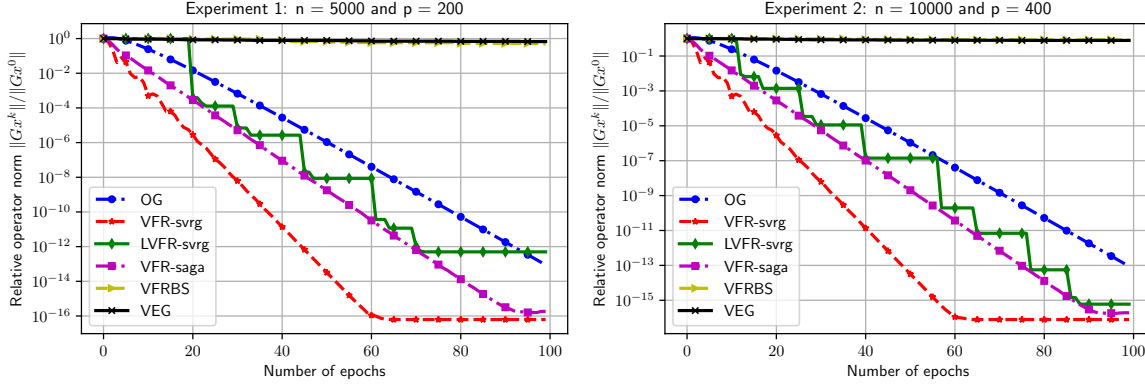


Figure 8: Performance of 6 algorithms for a large $b = \lfloor 0.1n \rfloor$ and a small $\mathbf{p} = \frac{1}{n^{1/2}}$.

As we can observed from Figure 8, our methods highly outperform VFRBS and VEG, suggesting that these competitors require a larger probability to select the snap-shot point w^k for full-batch evaluation. This is certainly not suggested in their theoretical results.

E.1.3. THE CONSTRAINED CASE

Next, we choose $f(\theta) = \delta_{[-r, r]^{p_1}}(\theta)$ and $g(\beta) := \delta_{[-r, r]^{p_2}}(\beta)$ as the indicators of the ℓ_∞ -balls of radius $r = 5$, respectively. In this case, we implement three variants of (VFRBS): the double-loop (VFR-svrg), the loopless (LVFR-svrg), and the SAGA (VFR-saga) variants to solve (NI) and compare against 3 algorithms as in the unconstrained case. Using the same data generating procedure as in the unconstrained case, we obtain the results as shown in Figure 9 when $K = \mathbb{I}$.

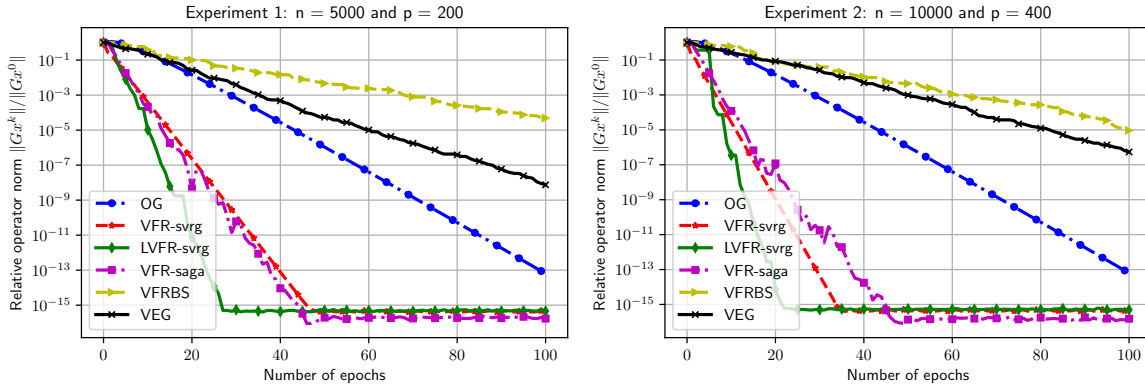


Figure 9: Comparison of 6 algorithms to solve constrained instances of (37) on 2 experiments when $K = \mathbb{I}$ (The average of 10 runs).

We see that the two SVRG variants of our (VFRBS): VFR-svrg and LVFR-svrg, as well as our VFR-saga variant remain working well compared to other methods. They superiorly outperform the three competitors.

Finally, we test our methods and their competitors for the case $K \neq \mathbb{I}$ as we done in Figure 5. Our results are plotted in Figure 10, where we observe a similar behavior as in Figure 5.

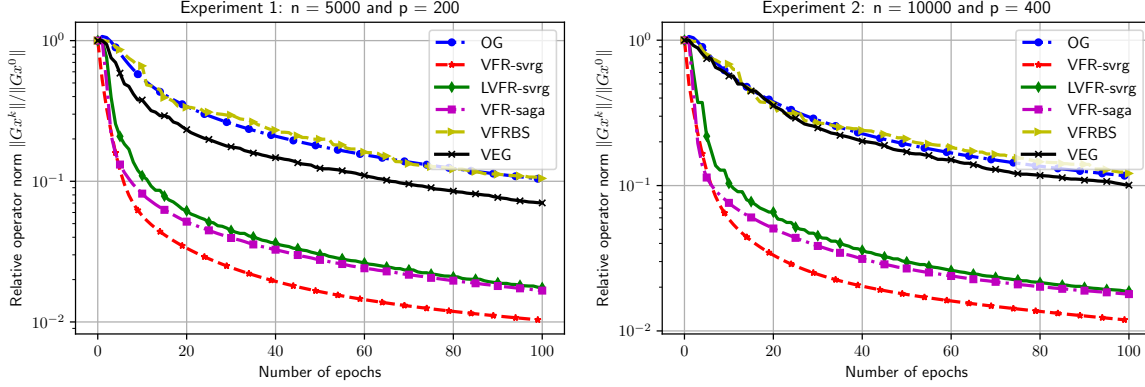


Figure 10: Comparison of 6 algorithms to solve constrained instances of (37) on 2 experiments when $K \neq \mathbb{I}$ (The average of 10 runs).

E.2. Nonconvex-Nonconcave Quadratic Minimax Problems

We recall the nonconvex-nonconcave quadratic minimax optimization problem (15) in this subsection:

$$\min_{u \in \mathbb{R}^{p_1}} \max_{v \in \mathbb{R}^{p_2}} \left\{ \mathcal{L}(u, v) := \varphi(u) + \frac{1}{n} \sum_{i=1}^n [u^T A_i u + u^T L_i v - v^T B_i v + b_i^T u - c_i^T v] - \psi(v) \right\}, \quad (38)$$

where $A_i \in \mathbb{R}^{p_1 \times p_1}$ and $B_i \in \mathbb{R}^{p_2 \times p_2}$ are symmetric matrices, $L_i \in \mathbb{R}^{p_1 \times p_2}$, $b_i \in \mathbb{R}^{p_1}$, $c_i \in \mathbb{R}^{p_2}$, and $\varphi = \delta_{\Delta_{p_1}}$ and $\psi = \delta_{\Delta_{p_2}}$ are the indicator of standard simplexes in \mathbb{R}^{p_1} and \mathbb{R}^{p_2} , respectively.

Let us first define $x := [u, v] \in \mathbb{R}^p$ as the concatenation of the primal and dual variables u and v , where $p := p_1 + p_2$. Next, we define

$$G_i x = \mathbf{G}_i x + \mathbf{g}_i := \begin{bmatrix} A_i & L_i \\ -L_i & B_i \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} b_i \\ c_i \end{bmatrix} = \begin{bmatrix} A_i u + L_i v + b_i \\ -L_i u + B_i v + c_i \end{bmatrix}, \quad \text{and} \quad T := \begin{bmatrix} \partial \varphi \\ \partial \psi \end{bmatrix}.$$

Then, we denote $\mathbf{G}_i := \begin{bmatrix} A_i & L_i \\ -L_i & B_i \end{bmatrix}$, and $\mathbf{g}_i := \begin{bmatrix} b_i \\ c_i \end{bmatrix}$. Clearly, $G_i(\cdot)$ is an affine mapping from \mathbb{R}^p to \mathbb{R}^p , but \mathbf{G}_i is nonsymmetric. Let $Gx := \frac{1}{n} \sum_{i=1}^n G_i x = (\frac{1}{n} \sum_{i=1}^n \mathbf{G}_i) x + \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i = \mathbf{G}x + \mathbf{g}$, where $\mathbf{G} := \frac{1}{n} \sum_{i=1}^n \mathbf{G}_i$ and $\mathbf{g} := \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i$. Then, the optimality condition of (38) becomes $0 \in Gx + Tx$, which is exactly in the form (NI). Clearly, if A_i and/or B_i are not positive semidefinite, then (38) possibly covers nonconvex-nonconcave minimax optimization instances.

E.2.1. THE UNCONSTRAINED CASE

We consider the case $\varphi = 0$ and $\psi = 0$, leading to an unconstrained setting of (38), i.e. $T = 0$ as considered in (15) of the main text. Hence, the optimality condition of (38) reduces to $Gx = 0$, which is of the form (NE).

(a) **How to generate data?** To run our experiments, we generate synthetic data as follows. First, we fix the dimensions p_1 and p_2 and the number of components n . We generate $A_i = Q_i D_i Q_i^T$ for a given orthonormal matrix Q_i and a diagonal matrix $D_i = \text{diag}(D_i^1, \dots, D_i^{p_1})$, where its elements are generated from standard normal distribution and clipped its negative entries as $\max\{D_i^j, \varepsilon\}$ for $j = 1, \dots, p_1$ and $\varepsilon := -0.1$. This choice of A_i guarantees that A_i is symmetric, but possibly not positive semidefinite. The matrix B_i is also generated by the same way. The pay-off matrix L_i is an $p_1 \times p_2$ matrix, which is also generated from the standard normal distribution for all $i \in [n]$. The vectors b_i and c_i are generated from the standard normal distribution for $i \in [n]$. With this data generating procedure, \mathbf{G}_i is not symmetric and possibly not positive semidefinite.

(b) **Algorithms.** We again test 6 algorithms: two variants (double-loop SVRG – VFR-svrg) and (loopless SVRG – LVFR-svrg) of (VFR), our (VFR) with SAGA estimator (VFR-saga), VFRBS from (Alacaoglu et al., 2023), VEG from (Alacaoglu & Malitsky, 2022), and OG (the standard optimistic gradient method), e.g., from (Daskalakis et al., 2018).

(c) **The details of Subsection 5.1 in Section 5.** First, we provide the details of Subsection 5.1 in Section 5. The purpose of this example is to verify our theoretical results stated in Corollaries 3.2 and 3.3.

For the SVRG estimator, let us first choose $\gamma := 0.75$, $b := \lfloor n^{2/3} \rfloor$, and $\mathbf{p} := \frac{1}{n^{1/3}}$ as suggested by Corollary 3.2. Then, we can directly compute $\eta := \frac{1}{L\sqrt{M}}$, where $\Lambda := \frac{6.25(2-3\mathbf{p})+4.125\mathbf{p}^2}{b\mathbf{p}^2}$ and $M = 2.375 + \frac{11}{3}\Lambda$. Clearly, if $n = 5000$, then $\eta = \frac{0.146153}{L}$. Alternatively, if $n = 10000$, then $\eta = \frac{0.148934}{L}$. These learning rates are used in our experiments plotted in Figure 1.

Similarly, for the SAGA estimator, we also choose $\gamma := 0.75$ and $b := \lfloor n^{2/3} \rfloor$. In this case, by Corollary 3.3, we can also directly compute $\eta := \frac{1}{L\sqrt{M}}$. If $n = 5000$, then $\eta = \frac{0.146153}{L}$. Alternatively, if $n = 10000$, then $\eta = \frac{0.145693}{L}$. These learning rates are used in VFR-saga.

Note that since the theoretical value of \mathbf{p} in VFRBS and VEG is too small, we instead choose $\mathbf{p} := \frac{1}{n^{1/3}}$ and also $b := \lfloor n^{2/3} \rfloor$ as in our methods. Then, we compute the learning rate η of these methods based on the formula given in (Alacaoglu et al., 2023) for VFRBS and (Alacaoglu & Malitsky, 2022) for VEG, respectively.

(d) **Results for a different set of parameters.** Unlike Subsection 5.1 in the main text, we choose the parameters for these algorithms as in Subsection E.1. The 6 algorithms are run on 2 experiments. The first experiment is with $n = 5000$ and $p_1 = p_2 = 50$, while the second one is with $n = 10000$ and $p_1 = p_2 = 100$. These experiments are run 10 times, corresponding to 10 problem instances, and the average results are reported in Figure 11 in terms of the relative operator norm $\|Gx^k\|/\|Gx^0\|$ against the number of epochs.

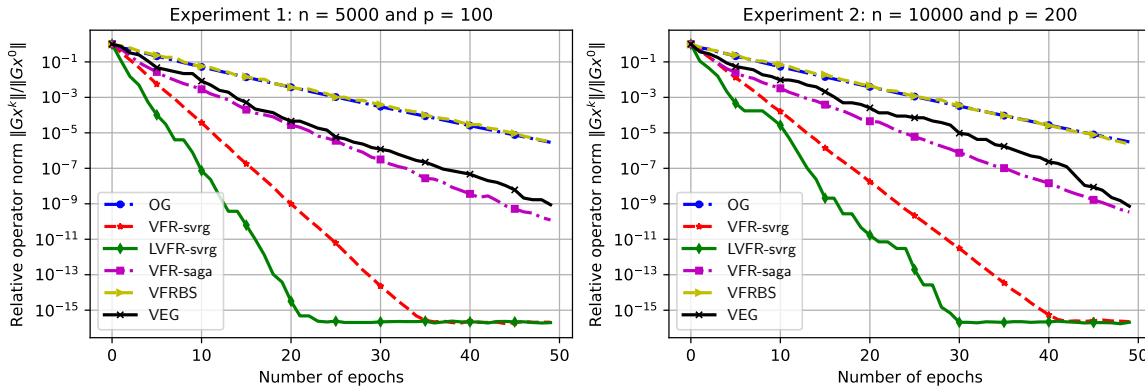


Figure 11: The performance of 6 algorithms to solve the unconstrained case of (38) on 2 experiments (The average of 10 runs).

Clearly, under this configuration, both SVRG variants of our methods work well and significantly outperform other competitors. The loopless SVRG variant (VFR-svrg) of (VFR) seems to work best, while our VFR-saga has a similar performance as VEG. We also see that VFRBS has a similar performance as OG.

To improve the performance of these competitors, especially, VFRBS and VEG, one can tune their parameters as in Subsection E.1, where the probability \mathbf{p} of updating the snapshot point w^k is increased. However, with such a choice of \mathbf{p} , its value is often greater or equal to 0.5, making these methods to be closed to deterministic variants. Hence, their theoretical complexity bounds are no longer improved over the deterministic counterparts.

E.2.2. THE CONSTRAINED CASE

We conduct two more experiments for the constrained case of (38) as in the main text when $u \in \Delta_{p_1}$ and $v \in \Delta_{p_2}$, where $\Delta_p := \{u \in \mathbb{R}_+^p : \sum_{i=1}^p u_i = 1\}$ is the standard simplex in \mathbb{R}^p .

We run 6 algorithms for solving the constrained case of (38) using the same parameters as Subsection 5.1, but with larger problems. We report the relative norm of the FBS residual $\|\mathcal{G}_\eta x^k\|/\|\mathcal{G}_\eta x^0\|$ against the number of epochs. The results are revealed in Figure 12 for two datasets $(p, n) = (500, 5000)$ and $(p, n) = (300, 10000)$.

With these two additional experiments, both SVRG variants of our method (VFRBS) again work well and significantly outperform other competitors. The loopless SVRG variant (VFR-svrg) of (VFRBS) tends to work best, while our VFR-saga has a relatively similar performance as VEG. We also see that VFRBS has a similar performance trend as OG,

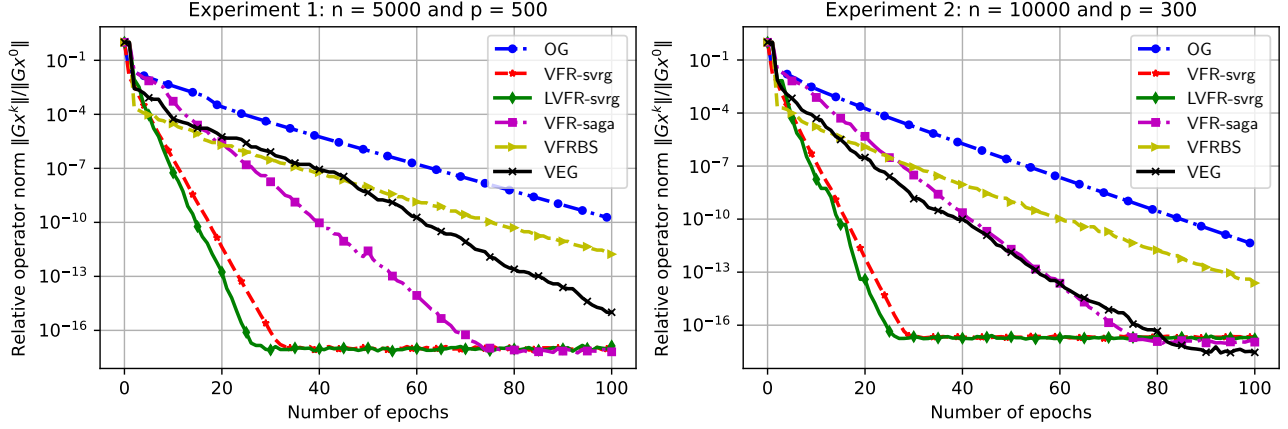


Figure 12: The performance of 6 algorithms to solve the constrained case of (38) on 2 experiments (The average of 10 runs).

but works better.

E.3. The ℓ_1 -Regularized Logistic Regression with Ambiguous Features

This Supp. Doc. provides the details of **Subsection 5.2** in Section 5 in the main text.

(a) **Model.** We consider a standard regularized logistic regression model associated with a given dataset $\{(\hat{X}_i, y_i)\}_{i=1}^N$, where \hat{X}_i is an i.i.d. sample of a feature vector and $y_i \in \{0, 1\}$ is the associated label of \hat{X}_i . Unfortunately, \hat{X}_i is ambiguous, i.e. it belongs to one of m possible examples $\{X_{ij}\}_{j=1}^m$. Since we do not know \hat{X}_i to evaluate the loss, we consider the worst-case loss $f_i(w) := \max_{1 \leq j \leq m} \ell(\langle X_{ij}, w \rangle, y_i)$ computed from m examples, where $\ell(\tau, s) := \log(1 + \exp(\tau)) - s\tau$ is the standard logistic loss.

Using the fact that $\max_{1 \leq j \leq m} \ell_j(\cdot) = \max_{z \in \Delta_m} \sum_{j=1}^m z_j \ell_j(\cdot)$, where Δ_m is the standard simplex in \mathbb{R}^m , we can model this regularized logistic regression into the following minimax problem:

$$\min_{w \in \mathbb{R}^d} \max_{z \in \mathbb{R}^m} \left\{ \mathcal{L}(w, z) := \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m z_j \ell(\langle X_{ij}, w \rangle, y_i) + \tau R(w) - \delta_{\Delta_m}(z) \right\}, \quad (39)$$

where $\ell(\tau, s) := \log(1 + \exp(\tau)) - s\tau$ is the standard logistic loss, $R(w) := \|w\|_1$ is an ℓ_1 -norm regularizer, $\tau > 0$ is a regularization parameter, and δ_{Δ_m} is the indicator of Δ_m that handles the constraint $z \in \Delta_m$. This problem is exactly the one stated in (16) of the main text.

First, let us denote $x := [w; z] \in \mathbb{R}^p$ as the concatenation of w and z with $p = d + m$, and

$$G_i x := \begin{bmatrix} \sum_{j=1}^m z_j \ell'(\langle X_{ij}, w \rangle, y_i) X_{ij} \\ -\ell(\langle X_{i1}, w \rangle, y_i) \\ \vdots \\ -\ell(\langle X_{im}, w \rangle, y_i) \end{bmatrix} \quad \text{and} \quad T x := \begin{bmatrix} \tau \partial R(w) \\ \partial \delta_{\Delta_m}(z) \end{bmatrix},$$

where $\ell'(\tau, s) = \frac{\exp(\tau)}{1 + \exp(\tau)} - s$. Then, the optimality condition of (39) can be written as (NI): $0 \in Gx + Tx$, where $Gx := \frac{1}{N} \sum_{i=1}^N G_i x$.

(b) **Input data.** We test our algorithms and their competitors on two real datasets: a9a (134 features and 3561 samples) and w8a (311 features and 45546 samples) downloaded from LIBSVM (Chang & Lin, 2011). For a given nominal dataset $\{(\hat{X}_i, y_i)\}_{i=1}^n$, we first normalize the feature vector \hat{X}_i such that its column norm is one, and then add a column of all ones to address the bias term. To generate ambiguous features, we take the nominal feature vector \hat{X}_i and add a random noise generated from a normal distribution of zero mean and variance of $\sigma = 0.5$. In our test, we choose $\tau := 10^{-3}$ and $m := 10$ for all the experiments.

(c) **Algorithms.** As before, we implement 3 variants of our method (VFRBS): VFR-svrg, LVFR-svrg, and VFR-saga

to solve (39). We also compare them with OG, VFRBS, and VEG. We choose $x^0 := 0.5 \cdot \text{ones}(p)$ in all experiments. We run all the algorithms for 100 epochs and report the relative FBS residual norm $\|\mathcal{G}_\eta x^k\|/\|\mathcal{G}_\eta x^0\|$ against the epochs.

(d) **Parameters.** Since it is very difficult to estimate the Lipschitz constant L of G , we are unable to set a correct learning rate η in the underlying algorithms. We instead compute an estimation $\hat{L} := \|\hat{X}\|$, and then set $\eta := \frac{\omega}{\hat{L}}$, by tuning ω for each algorithm. More specifically, after tuning, we obtain the following configuration.

- For the three variants of (VFRBS): VFR-svrg, LVFR-svrg, and VFR-saga, we set $\eta = \frac{25}{\hat{L}}$ for a9a and $\eta = \frac{50}{\hat{L}}$ for w8a.
- For OG, we set $\eta = \frac{50}{\hat{L}}$ for a9a and $\eta = \frac{100}{\hat{L}}$ for w8a.
- For VFRBS, we choose $\eta = \frac{47.5(1-\sqrt{1-\mathbf{p}})}{2\hat{L}}$ for a9a and $\eta = \frac{95(1-\sqrt{1-\mathbf{p}})}{2\hat{L}}$ for w8a.
- For VEG, we select $\eta = \frac{47.5\sqrt{1-\alpha}}{\hat{L}}$ for a9a and $\eta = \frac{95\sqrt{1-\alpha}}{\hat{L}}$ for w8a with $\alpha := 1 - \mathbf{p}$.

We still choose the mini-batch size b and the probability \mathbf{p} of updating the snapshot point w^k in SVRG variants as $b = \lfloor 0.5n^{2/3} \rfloor$ and $\mathbf{p} = n^{-1/3}$, respectively for all the algorithms.

We conduct two more experiments using the well-known MNIST dataset ($n = 70000$ and $p = 780$) where we want to classify the even and odd numbers into two different classes, respectively. We use the same parameter selection as in the experiment with the a9a dataset. In the first experiment, we choose $m = 10$ and the variance of noise $\sigma = 0.5$. In the second experiment, we choose $m = 20$ and the variance of noise $\sigma = 0.25$. We only run 5 algorithms and leave out the VFRBS method since we have not managed to find the parameters that make it work stably. The result of this experiment is shown in Figure 13.

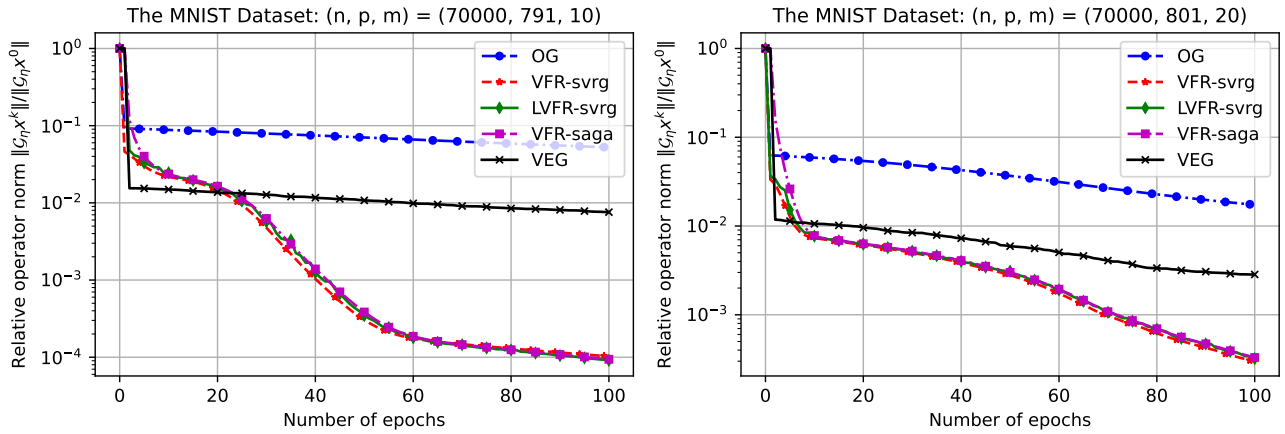


Figure 13: Comparison of 6 algorithms to solve (16) on the real dataset: MNIST.

We can see from Figure 13 again that three variants VFR-svrg, LVFR-svrg, and VFR-saga have similar performance and are much better than their two competitors. Here, VEG is better than OG, but both methods are slower than ours.

References

- Adly, S. and Attouch, H. First-order inertial algorithms involving dry friction damping. *Math. Program.*, pp. 1–41, 2021.
- Alacaoglu, A. and Malitsky, Y. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 778–816. PMLR, 2022.
- Alacaoglu, A., Böhm, A., and Malitsky, Y. Beyond the golden ratio for variational inequality algorithms. *J. Mach. Learn. Res.*, 24(172):1–33, 2023.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.
- Attouch, H. and Cabot, A. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Math. Program.*, 184(1):243–287, 2020.

- Bauschke, H. H., Moursi, W. M., and Wang, X. Generalized monotone operators and their averaged resolvents. *Math. Program.*, pp. 1–20, 2020.
- Beznosikov, A., Gorbunov, E., Berard, H., and Loizou, N. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 172–235. PMLR, 2023.
- Böhm, A. Solving nonconvex-nonconcave min-max problems exhibiting weak Minty solutions. *Transactions on Machine Learning Research*, 2022.
- Boş, R. I., Mertikopoulos, P., Staudigl, M., and Vuong, P. T. Minibatch forward-backward-forward methods for solving stochastic variational inequalities. *Stochastic Systems*, 11(2):112–139, 2021.
- Cai, X., Song, C., Guzmán, C., and Diakonikolas, J. A stochastic Halpern iteration with variance reduction for stochastic monotone inclusion problems. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2022)*, 2022. URL <https://openreview.net/forum?id=BRZos-8TpCf>.
- Cai, X., Alacaoglu, A., and Diakonikolas, J. Variance reduced halpern iteration for finite-sum monotone inclusions. In *The 12th International Conference on Learning Representations (ICLR)*, pp. 1–33, 2024.
- Chakrabarti, D., Diakonikolas, J., and Kroer, C. Block-coordinate methods and restarting for solving extensive-form games. *Advances in Neural Information Processing Systems*, 36, 2024.
- Chang, C.-C. and Lin, C.-J. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- Combettes, P. L. and Eckstein, J. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Math. Program.*, 168(1):645–672, 2018.
- Combettes, P. L. and Pesquet, J.-C. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.*, 25(2):1221–1248, 2015.
- Cui, S. and Shanbhag, U. 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.
- Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, pp. 15210–15219, 2019.
- Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with Optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.
- Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1646–1654, 2014.
- Diakonikolas, J., Daskalakis, C., and Jordan, M. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.
- Driggs, D., Liang, J., and Schönlieb, C.-B. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022.
- Gorbunov, E., Berard, H., Gidel, G., and Loizou, N. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pp. 7865–7901. PMLR, 2022a.
- Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: $\mathcal{O}(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022b.
- Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Math. Program.*, 188(1):135–192, 2021.

- Hamedani, E. Y., Jalilzadeh, A., Aybat, N. S., and Shanbhag, U. V. Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118*, 2018.
- Hanzely, F., Mishchenko, K., and Richtárik, P. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pp. 2082–2093, 2018.
- Horváth, S., Kovalev, D., Mishchenko, K., Richtárik, P., and Stich, S. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, 38(1):91–106, 2023.
- Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pp. 6938–6948, 2019.
- Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.*, 27(2):686–724, 2017.
- Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.
- Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- Kannan, A. and Shanbhag, U. V. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Comput. Optim. Appl.*, 74(3):779–820, 2019.
- Konnov, I. *Combined relaxation methods for variational inequalities*. Springer-Verlag, 2001.
- Kotsalis, G., Lan, G., and Li, T. Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation. *SIAM J. Optim.*, 32(3):2041–2073, 2022.
- Li, C. J., Yu, Y., Loizou, N., Gidel, G., Ma, Y., Roux, N. L., and Jordan, M. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pp. 9793–9826. PMLR, 2022.
- Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. 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.
- Luo, Y. and Tran-Dinh, Q. Extragradient-type methods for co-monotone root-finding problems. (*UNC-STOR Technical Report*), 2022.
- Malitsky, Y. and Tam, M. K. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM J. Optim.*, 30(2):1451–1472, 2020.
- Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.
- Noor, M. A. Extragradient methods for pseudomonotone variational inequalities. *J. Optim. Theory Appl.*, 117(3):475–488, 2003.
- Noor, M. A. and Al-Said, E. Wiener–Hopf equations technique for quasimonotone variational inequalities. *J. Optim. Theory Appl.*, 103:705–714, 1999.
- Peng, Z., Xu, Y., Yan, M., and Yin, W. ARock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM J. Scientific Comput.*, 38(5):2851–2879, 2016.
- Pethick, T., Fercoq, O., Latafat, P., Patrinos, P., and Cevher, V. Solving stochastic weak Minty variational inequalities without increasing batch size. In *Proceedings of International Conference on Learning Representations (ICLR)*, pp. 1–34, 2023.
- Pham, H. N., Nguyen, M. L., Phan, T. D., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *J. Mach. Learn. Res.*, 21:1–48, 2020.

- Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- Rockafellar, R. and Wets, R. J.-B. *Variational Analysis*. Springer-Verlag, 1997.
- Song, C. and Diakonikolas, J. Cyclic coordinate dual averaging with extrapolation. *SIAM J. Optim.*, 33(4):2935–2961, 2023.
- Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020.
- Tran-Dinh, Q. Sublinear Convergence Rates of Extragradient-Type Methods: A Survey on Classical and Recent Developments. *arXiv preprint arXiv:2303.17192*, 2023.
- Tran-Dinh, Q. and Luo, Y. Randomized Block-Coordinate Optimistic Gradient Algorithms for Root-Finding Problems. *Math. Oper. Res.*, in press, 2025.
- Tran-Dinh, Q. and Nguyen-Trung, N. Variance-Reduced Accelerated Fixed-Point-Based Methods for Generalized Equations: Better Convergence Guarantees. *Tech. Report – UNC-STOR*, 2025a.
- Tran-Dinh, Q. and Nguyen-Trung, N. Accelerated extragradient-type methods—part 2: Generalization and sublinear convergence rates under co-hypomonotonicity. *arXiv preprint arXiv:2501.04585*, 2025b.
- Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. *Math. Program.*, 191:1005–1071, 2022.
- Tu, V. P. On the weak convergence of the extragradient method for solving pseudo-monotone variational inequalities. *J. Optim. Theory Appl.*, 176(2):399–409, 2018.
- Yang, J., Kiyavash, N., and He, N. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.
- Yousefian, F., Nedić, A., and Shanbhag, U. V. 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.
- Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3925–3936. Curran Associates Inc., 2018.