

Supplementary Document:

VARIANCE-REDUCED FORWARD-REFLECTED ALGORITHMS FOR 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 A 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.

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 (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)). 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 (Bot 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 of 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.

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., 2023; 2022), co-coercive structures in (Beznosikov et al., 2023), and bilinear game models in (Li et al., 2022).

Among many existing works, perhaps, (Cai et al., 2023) 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., 2023) 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. (2023) 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. (2023) 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 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. (2023) suggests to use the variance-reduced FRBS method in Alacaoglu et al. (2022) to approximate this resolvent, leading to a double loop algorithm. 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. (2023). 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. (2023).

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 (2023); 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.

B THE PROOF OF TECHNICAL RESULTS IN SECTION 2

This supplementary section provides the full proof of Lemma 2.1 and Lemma 2.2.

Further discussion of FR operator. Let us recall our operator S_γ^k defined by (FRO) 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.

Other possible stochastic estimators for S_γ^k . One natural idea to construct an unbiased estimator for S_γ^k is to use an increasing mini-batch stochastic estimator as $\tilde{S}_\gamma^k := \frac{1}{b_k} \sum_{i \in B_k} [G_i x^k - \gamma G_i x^{k-1}]$, where B_k is an increasing mini-batch in $[n]$, with $b_k := |B_k| \geq \frac{b_{k-1}}{1-\rho_k} \geq b_{k-1}$, 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 (I) 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 2.1: LOOPLESS-SVRG ESTIMATOR

Let us further expand Lemma 2.1 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 (FRO) and \tilde{S}_γ^k be generated by (L-SVRG). We consider the following quantity:

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

Then, we have

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

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

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

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

Here, $\textcircled{1}$ holds due to the i.i.d. property of \mathcal{B}_k , and $\textcircled{2}$ 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

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

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 $\frac{2(1+c)(1-\mathbf{p})}{c} + \mathbf{p} = \frac{4-6\mathbf{p}+3\mathbf{p}^2}{\mathbf{p}}$. Hence, we obtain

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

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

B.2 PROOF OF LEMMA 2.2: SAGA ESTIMATOR

Similarly, we also further expand Lemma 2.2 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 (FRO) 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

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

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}^k , we have

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

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

Next, let $X_i := G_i x^k - \gamma G_i x^{k-1} - (1-\gamma)\hat{G}_i^k$ 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

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

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

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

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

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

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

C CONVERGENCE ANALYSIS OF VFR FOR (NE): TECHNICAL PROOFS

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

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

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 L.3 and L.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*

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

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

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

Second, it is obvious to show that

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

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

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

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

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

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 3.1 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 3.1 from this result by fixing $\gamma = \frac{3}{4}$.

Corollary C.1. Suppose that Assumptions L.1, L.3 and L.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

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

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:

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

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 2.1, 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 3.1. 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 3.2 for the case $\gamma \in (\frac{1}{2}, 1)$ and then derive the oracle complexity of Corollary 3.2 from this result by fixing $\gamma := \frac{3}{4}$.

Corollary C.2. Suppose that Assumptions [I.1](#), [I.3](#) and [I.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

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

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:

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

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

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 $K := \left\lfloor \Gamma \cdot \frac{L^2 R_0^2 n^2}{b^3 \epsilon^2} \right\rfloor$, 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

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

If we choose $b := \lfloor n^{2/3} \rfloor$, then $\mathcal{T}_c = \left\lfloor \frac{3\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \right\rfloor$. 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 $\mathcal{T}_{G_i} := n + \left\lfloor \frac{3\Gamma L^2 R_0^2 n^{2/3}}{\epsilon^2} \right\rfloor$. \square

Proof of Corollary [3.2](#) 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.2](#), we obtain our conclusions in Corollary [3.2](#). Moreover, [\(29\)](#) reduces to $\frac{1}{K+1} \sum_{k=0}^K \mathbb{E}[\|Gx^k\|^2] \leq \frac{32(1+0.494^2)L^2 R_0^2}{3 \cdot 0.1494^2 b^2 (K+1)} \leq \frac{489 \cdot L^2 R_0^2}{b^2 (K+1)}$ 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 (FRO). 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

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

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

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

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

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

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

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

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

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

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 4.1 and 4.2 in the main text. Unlike Corollaries 3.1 and 3.2 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 4.1 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 $\mathcal{T}_{G_i} = n + \left\lceil \frac{4\Gamma n^{2/3} L^2 \hat{R}_0^2}{\epsilon^2} \right\rceil$. Moreover, at each iteration, we need one evaluation of $J_{\gamma\eta T}$. Therefore, the total evaluations of $J_{\gamma\eta T}$ is $\mathcal{T}_T := K = \left\lceil \Gamma \cdot \frac{L^2 \hat{R}_0^2}{b\mathbf{p}^2 \epsilon^2} \right\rceil = \left\lceil \Gamma \cdot \frac{L^2 \hat{R}_0^2}{\epsilon^2} \right\rceil$. \square

Proof of Corollary 4.2 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 $\mathcal{T}_{G_i} :=$

$n + \lfloor \frac{3\Gamma L^2 \hat{R}_0^2 n^{2/3}}{\epsilon^2} \rfloor$ evaluations of G_i . Moreover, since each iteration, it requires one evaluation of $J_{\gamma\eta T}$, we need $\mathcal{T}_T := K = \lfloor \Gamma \cdot \frac{L^2 \hat{R}_0^2}{\epsilon^2} \rfloor$ evaluations of $J_{\gamma\eta T}$. \square

Remark D.1. 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. (2022)), and the variance-reduced extragradient (VEG) in (Alacaoglu & Malitsky, 2021).

(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. (2022)), we choose its learning rate $\eta := \frac{0.95(1-\sqrt{1-\mathbf{p}})}{2L}$ as suggested by (Alacaoglu et al. (2022)). 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. (2022)), typically $\mathbf{p} = \mathcal{O}(1/n)$.

For the variance reduction extragradient method (VEG) in (Alacaoglu & Malitsky (2021)), 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 3.

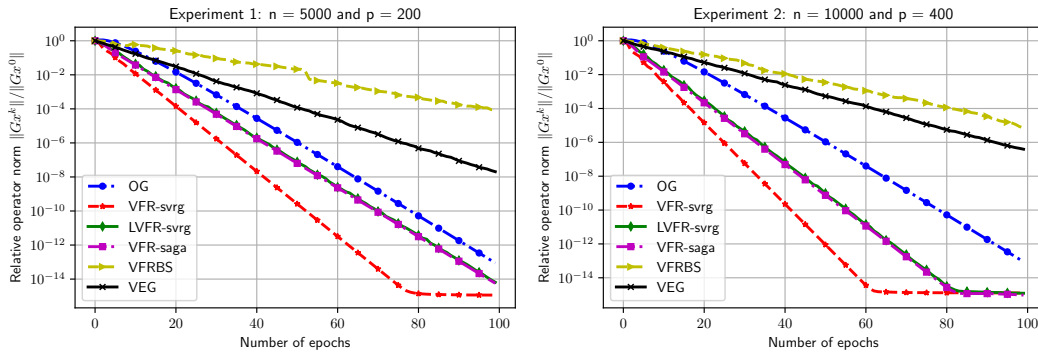


Figure 3: Performance of 6 algorithms to solve (37) on 2 experiments when $K = \mathbb{I}$.

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 3 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 4.

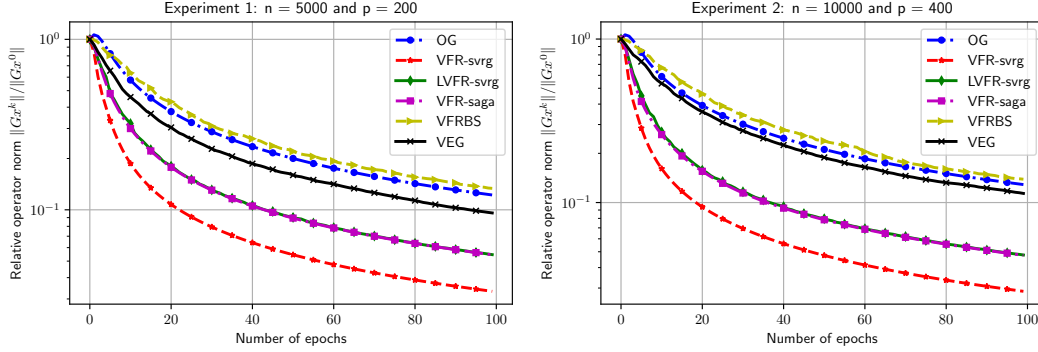


Figure 4: Performance of 6 algorithms to solve (37) on 2 experiments when $K \neq \mathbb{I}$.

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 5 reveals the performance of these algorithms when we increase the mini-batch size b to a larger value $b = \lfloor 0.1n \rfloor$, while keeping the probability $\mathbf{p} = \frac{1}{n^{1/3}}$ unchanged.

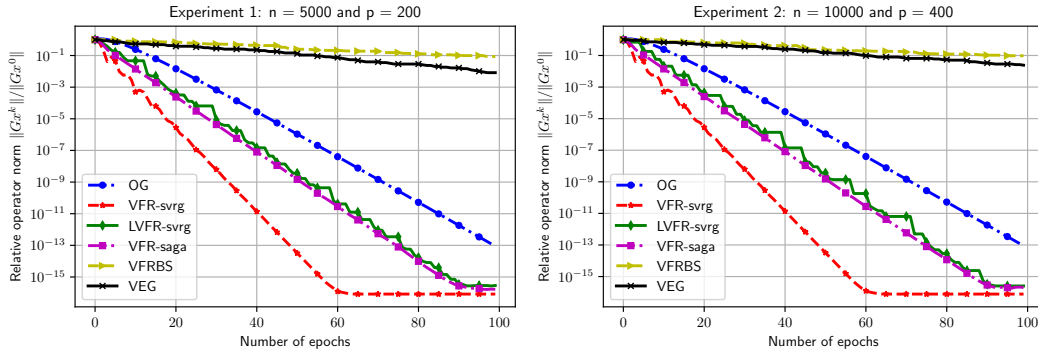


Figure 5: Performance of 6 algorithms for a large $b = \lfloor 0.1n \rfloor$ and a unchanged $\mathbf{p} = \frac{1}{n^{1/3}}$.

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 6.

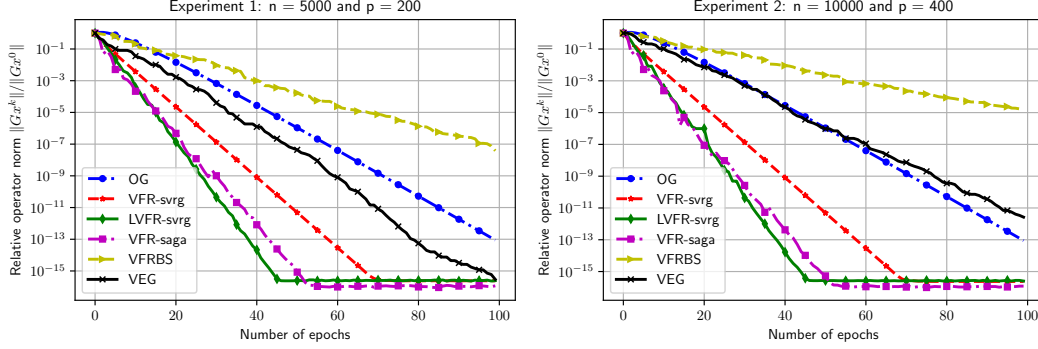


Figure 6: Performance of 6 algorithms for a medium $b = \lfloor 0.05n \rfloor$ and larger $\mathbf{p} = \frac{1}{n^{1/4}}$.

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 7.

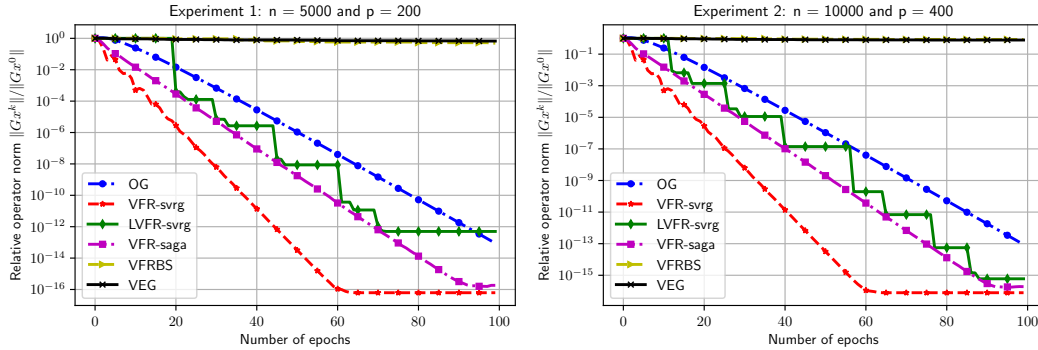


Figure 7: Performance of 6 algorithms for a large $b = \lfloor 0.1n \rfloor$ and a small $\mathbf{p} = \frac{1}{n^{1/2}}$.

As we can observed from Figure 7 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 8 when $K = \mathbb{I}$.

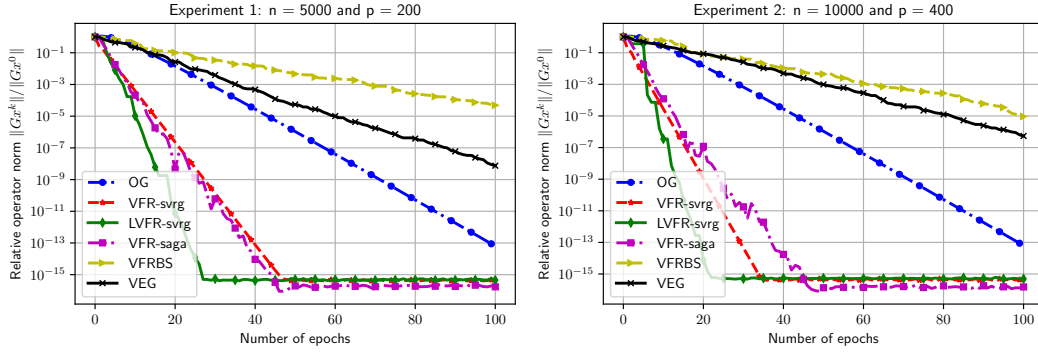


Figure 8: Comparison of 6 algorithms to solve constrained instances of (37) on 2 experiments when $K = \mathbb{I}$ (The average of 10 runs).

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 4. Our results are plotted in Figure 9, where we observe a similar behavior as in Figure 4.

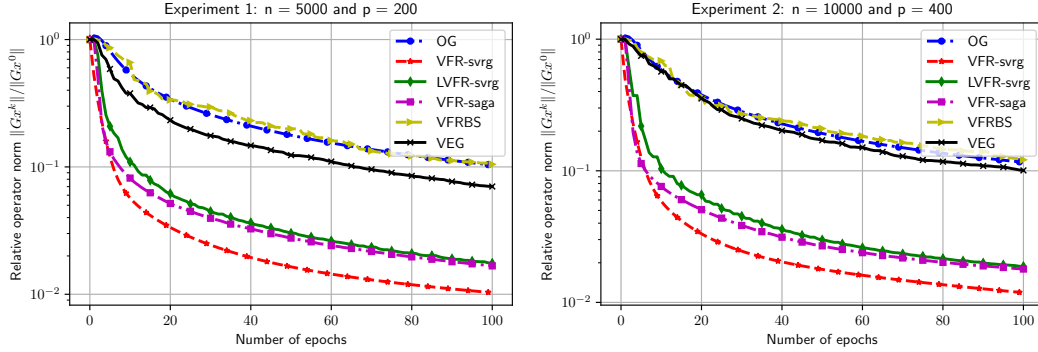


Figure 9: Comparison of 6 algorithms to solve constrained instances of (37) on 2 experiments when $K \neq \mathbb{I}$ (The average of 10 runs).

E.2 NONCONVEX-NONCONCAVE QUADRATIC MINIMAX PROBLEMS

We extend the nonconvex-nonconcave quadratic minimax optimization problem from the unconstrained setting (15) to the following constrained setting:

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

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 $f = \delta_{\Delta_{p_1}}$ and $g = \delta_{\Delta_{p_2}}$ 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

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

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 (NE). 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 $f = 0$ and $g = 0$, 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. (2022), VEG from Alacaoglu & Malitsky (2021), and OG (the standard optimistic gradient method), e.g., from Daskalakis et al. (2018).

(c) **The details of Example 1 in Section 5.** First, we provide the details of Example 1 in Section 5. The purpose of this example is to verify our theoretical results stated in Corollaries 3.1 and 3.2.

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 3.1. 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 $M = 2.375 + \frac{1}{3}\Lambda$. 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 3.2, 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. (2022) for VFRBS and Alacaoglu & Malitsky (2021) for VEG, respectively.

(d) **Results for a different set of parameters.** Unlike Example 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 10 in terms of the relative operator norm $\|Gx^k\|/\|Gx^0\|$ against the number of epochs.

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

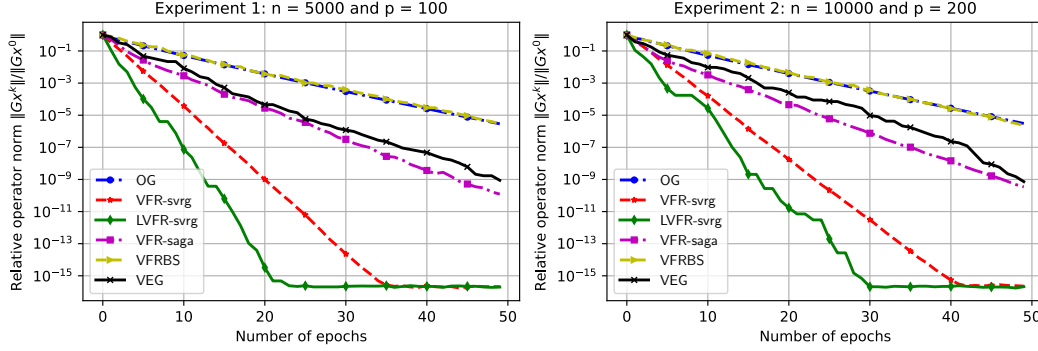


Figure 10: The performance of 6 algorithms to solve the unconstrained case of (38) on 2 experiments (The average of 10 runs).

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 now adding two simplex constraints $u \in \Delta_{p_1}$ and $v \in \Delta_{p_2}$ to (38), 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 $f(u) := \delta_{\Delta_{p_1}}(u)$ and $g(v) := \delta_{\Delta_{p_2}}(v)$ as the indicators of Δ_{p_1} and Δ_{p_2} , respectively. Under this setting, the optimality conditions of (38) becomes (N1), where $T := [\partial f, \partial g] = [\mathcal{N}_{\Delta_{p_1}}, \mathcal{N}_{\Delta_{p_2}}]$ with $\mathcal{N}_{\mathcal{X}}$ being the normal cone of \mathcal{X} . Hence, the resolvent $J_{\gamma\eta T}$ reduces to the projections on the simplex product $\Delta_{p_1} \times \Delta_{p_2}$.

Again, we run 6 algorithms for solving the constrained case of (38) using the same parameters as Subsection E.2.1. We report the relative norm of the FBS residual $\|G_\eta x^k\|/\|G_\eta x^0\|$ against the number of epochs. The results are revealed in Figure 11 for two datasets $(p, n) = (100, 5000)$ and $(p, n) = (200, 10000)$.

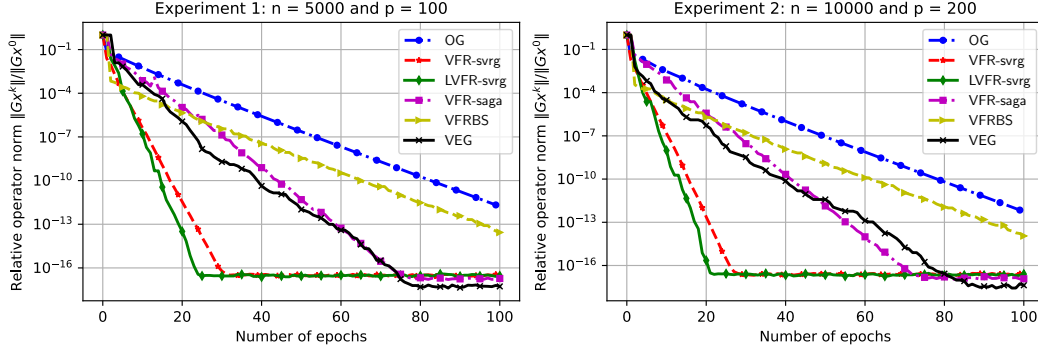


Figure 11: The performance of 6 algorithms to solve the constrained case of (38) on 2 experiments (The average of 10 runs).

Clearly, with these experiments, both SVRG variants of our method (VFRBS) work well and significantly outperform other competitors. The loopless 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.

E.3 THE ℓ_1 -REGULARIZED LOGISTIC REGRESSION WITH AMBIGUOUS FEATURES

This Supp. Doc. provides the details of **Example 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

$$G_i x := \begin{bmatrix} \sum_{j=1}^m z_j \ell'(\langle X_{ij}, w \rangle, y_i) X_{ij} \\ -\ell(\langle X_{i1}, w \rangle, y_i) \\ \dots \\ -\ell(\langle X_{im}, w \rangle, y_i) \end{bmatrix} \quad \text{and} \quad Tx := \begin{bmatrix} \tau \partial R(w) \\ \partial \delta_{\Delta_m}(z) \end{bmatrix},$$

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 $\|G_\eta x^k\| / \|G_\eta x^0\|$ 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.

REFERENCES

- S. Adly and H. Attouch. First-order inertial algorithms involving dry friction damping. *Math. Program.*, pp. 1–41, 2021.
- A. Alacaoglu and Y. Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

- A. Alacaoglu, A. Böhm, and Y. Malitsky. Beyond the golden ratio for variational inequality algorithms. *arXiv preprint arXiv:2212.13955*, 2022.
- M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.
- H. Attouch and A. Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Math. Program.*, 184(1):243–287, 2020.
- H. H. Bauschke, W. M. Moursi, and X. Wang. Generalized monotone operators and their averaged resolvents. *Math. Program.*, pp. 1–20, 2020.
- A. Beznosikov, E. Gorbunov, H. Berard, and N. Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 172–235. PMLR, 2023.
- A. Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak Minty solutions. *Transactions on Machine Learning Research*, 2022.
- R. I. Boş, P. Mertikopoulos, M. Staudigl, and P. T. Vuong. Minibatch forward-backward-forward methods for solving stochastic variational inequalities. *Stochastic Systems*, 11(2):112–139, 2021.
- X. Cai, C. Song, C. Guzmán, and J. Diakonikolas. A stochastic halpern iteration with variance reduction for stochastic monotone inclusion problems. *arXiv preprint arXiv:2203.09436*, 2022.
- X. Cai, A. Alacaoglu, and J. Diakonikolas. Variance reduced Halpern iteration for finite-sum monotone inclusions. *arXiv preprint arXiv:2310.02987*, 2023.
- D. Chakrabarti, J. Diakonikolas, and C. Kroer. Block-coordinate methods and restarting for solving extensive-form games. *Advances in Neural Information Processing Systems*, 36, 2024.
- C.-C. Chang and C.-J. Lin. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- P. L. Combettes and J. Eckstein. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Math. Program.*, 168(1):645–672, 2018.
- P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.*, 25(2):1221–1248, 2015.
- S. Cui and U.V. Shanbhag. On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *Set-Valued and Variational Analysis*, 29(2):453–499, 2021.
- C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with Optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.
- A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1646–1654, 2014.
- J. Diakonikolas, C. Daskalakis, and M. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.
- E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pp. 7865–7901. PMLR, 2022a.
- E. Gorbunov, N. Loizou, and G. Gidel. 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.
- R. M. Gower, P. Richtárik, and F. Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Math. Program.*, 188(1):135–192, 2021.

- E. Y. Hamedani, A. Jalilzadeh, N. S. Aybat, and U. V. Shanbhag. Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118*, 2018.
- F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pp. 2082–2093, 2018.
- S. Horváth, D. Kovalev, K. Mishchenko, P. Richtárik, and S. Stich. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, 38(1):91–106, 2023.
- Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pp. 6938–6948, 2019.
- A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.*, 27(2):686–724, 2017.
- Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.
- A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- A. Kannan and U. V. Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Comput. Optim. Appl.*, 74(3):779–820, 2019.
- I.V. Konnov. *Combined relaxation methods for variational inequalities*. Springer-Verlag, 2001.
- G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation. *SIAM J. Optim.*, 32(3):2041–2073, 2022.
- C. J. Li, Y. Yu, N. Loizou, G. Gidel, Y. Ma, N. Le Roux, and M. Jordan. 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.
- N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected cocoercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.
- Y. Luo and Q. Tran-Dinh. Extragradient-type methods for co-monotone root-finding problems. (*UNC-STOR Technical Report*), 2022.
- Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM J. Optim.*, 30(2):1451–1472, 2020.
- K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.
- M. A. Noor. Extragradient methods for pseudomonotone variational inequalities. *J. Optim. Theory Appl.*, 117(3):475–488, 2003.
- M. A. Noor and E.A. Al-Said. Wiener-Hopf equations technique for quasimonotone variational inequalities. *J. Optim. Theory Appl.*, 103:705–714, 1999.
- Z. Peng, Y. Xu, M. Yan, and W. Yin. ARock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM J. Scientific Comput.*, 38(5):2851–2879, 2016.
- T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher. Solving stochastic weak Minty variational inequalities without increasing batch size. *arXiv preprint arXiv:2302.09029*, 2023.
- H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

- 1998 R.T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Springer-Verlag, 1997.
- 1999
- 2000 C. Song and J. Diakonikolas. Cyclic coordinate dual averaging with extrapolation. *SIAM J. Optim.*, 33(4):2935–2961, 2023.
- 2001
- 2002 C. Song, Z. Zhou, Y. Zhou, Y. Jiang, and Y. Ma. Optimistic dual extrapolation for coherent non-
- 2003 monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:
- 2004 14303–14314, 2020.
- 2005
- 2006 Q. Tran-Dinh. Sublinear Convergence Rates of Extragradient-Type Methods: A Survey on Classical
- 2007 and Recent Developments. *arXiv preprint arXiv:2303.17192*, 2023.
- 2008 Q. Tran-Dinh and Y. Luo. Randomized block-coordinate optimistic gradient algorithms for root-
- 2009 finding problems. *arXiv preprint arXiv:2301.03113*, 2023.
- 2010
- 2011 V. Phan Tu. On the weak convergence of the extragradient method for solving pseudo-monotone
- 2012 variational inequalities. *J. Optim. Theory Appl.*, 176(2):399–409, 2018.
- 2013 J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class
- 2014 of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.
- 2015
- 2016 F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic mirror-prox algorithms for stochastic
- 2017 cartesian variational inequalities: Randomized block coordinate and optimal averaging schemes.
- 2018 *Set-Valued and Variational Analysis*, 26:789–819, 2018.
- 2019
- 2020
- 2021
- 2022
- 2023
- 2024
- 2025
- 2026
- 2027
- 2028
- 2029
- 2030
- 2031
- 2032
- 2033
- 2034
- 2035
- 2036
- 2037
- 2038
- 2039
- 2040
- 2041
- 2042
- 2043
- 2044
- 2045
- 2046
- 2047
- 2048
- 2049
- 2050
- 2051