# Stochastic Variance Reduction for Variational Inequality Methods

**Ahmet Alacaoglu** ALACAOGLU@WISC.EDU
*University of Wisconsin-Madison*

**Yura Malitsky** YURII.MALITSKYI@LIU.SE
*Linköping University*

**Editors:** Po-Ling Loh and Maxim Raginsky

## Abstract

We propose stochastic variance reduced algorithms for solving convex-concave saddle point problems, monotone variational inequalities, and monotone inclusions. Our framework applies to extragradient, forward-backward-forward, and forward-reflected-backward methods both in Euclidean and Bregman setups. All proposed methods converge in the same setting as their deterministic counterparts and they either match or improve the best-known complexities for solving structured min-max problems. Our results reinforce the correspondence between variance reduction in variational inequalities and minimization. We also illustrate the improvements of our approach with numerical evaluations on matrix games.

**Keywords:** Variational inequality, extragradient, stochastic methods, variance reduction, oracle complexity

## 1. Introduction

In this paper, we focus on solving variational inequalities (VI):

$$\text{find} \quad \mathbf{z}_* \in \mathcal{Z} \quad \text{such that} \quad \langle F(\mathbf{z}_*), \mathbf{z} - \mathbf{z}_* \rangle + g(\mathbf{z}) - g(\mathbf{z}_*) \geq 0, \quad \forall \mathbf{z} \in \mathcal{Z}, \tag{1}$$

where $F$ is a monotone operator and $g$ is a proper convex lower semicontinuous function. This formulation captures optimality conditions for minimization/saddle point problems, see (Facchinei and Pang, 2007, Sec. 1.4.1).

In the last decade there have been at least two surges of interest to VIs. Both were motivated by the need to solve min-max problems. The first surge came from the realization that many nonsmooth problems can be solved more efficiently if they are formulated as saddle point problems (Nesterov, 2005; Nemirovski, 2004; Chambolle and Pock, 2011; Esser et al., 2010). The second has been started by machine learning community, where solving nonconvex-nonconcave saddle point problems became of paramount importance (Gidel et al., 2019; Gemp and Mahadevan, 2018; Mertikopoulos et al., 2019). Additionally, VIs have applications in game theory, control theory, and differential equations, see (Facchinei and Pang, 2007).

A common structure encountered in min-max problems is that the operator $F$ can be written as a finite-sum: $F = F_1 + \cdots + F_N$, see App. D for concrete examples. Variance reduction techniques use this specific form to improve the complexity of deterministic methods in minimization. Existing results on variance reduction for saddle point problems show that these techniques improve the complexity for bilinear problems compared to deterministic methods. However, in general these methods require stronger assumptions to converge than the latter do (see Table 1). At the same

time, stochastic methods that have been shown to converge under only monotonicity do not have complexity advantages over the deterministic methods.

Such a dichotomy does not exist in minimization: variance reduction comes with no extra assumptions. This points out to a fundamental lack of understanding for its use in saddle point problems. Our work shows that there is indeed a natural correspondence between variance reduction in variational inequalities and minimization. In particular, we propose stochastic variants of extragradient (EG), forward-backward-forward (FBF), and forward-reflected-backward (FoRB) methods which converge under mere monotonicity. For the bilinear case our results match the best-known complexities, while for the nonbilinear, we do not require bounded domains as in the previous work and we improve the best-known complexity by a logarithmic factor, using simpler algorithms. Han et al. (2021) established the optimality of our algorithms with matching lower bounds, for solving (potentially nonbilinear) convex-concave min-max problems with finite sum form.

We also show application of our techniques for solving monotone inclusions and strongly monotone problems. Our results for monotone inclusions potentially improve the rate of deterministic methods (depending on the Lipschitz constants) and they seem to be the first such result in the literature. We illustrate practical benefits of our new algorithms by comparing with deterministic methods and an existing variance reduction scheme in App. E.

### 1.1. Related works

**Variational inequalities.** The standard choices for solving VIs have been methods such as extragradient (EG)/Mirror-Prox (MP) (Korpelevich, 1976; Nemirovski, 2004), forward-backward-forward (FBF) (Tseng, 2000), dual extrapolation (Nesterov, 2007) or reflected gradient/forward-reflected-backward (FoRB) (Malitsky, 2015; Malitsky and Tam, 2020)[1]. These methods differ in the number of operator calls and projections (or proximal operators) used each iteration, and consequently, can be preferable to one another in different settings. The standard convergence results for these algorithms include global iterates' convergence, complexity $\mathcal{O}(\varepsilon^{-1})$ for monotone problems and linear rate of convergence for strongly monotone problems.

**Variance reduction.** Variance reduction has revolutionized stochastic methods in optimization. This technique applies to finite sum minimization problem of the form $\min_{\mathbf{x}} \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x})$. Instead of using a random sample $\mathbf{g}_k = \nabla f_i(\mathbf{x}_k)$ as SGD does, variance reduction methods use

$$\mathbf{g}_k = \nabla f(\mathbf{w}_k) + \nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{w}_k). \tag{2}$$

A good choice of $\mathbf{w}_k$ decreases the "variance" $\mathbb{E}\,\|\mathbf{g_k} - \nabla f(\mathbf{x}_k)\|^2$ compared to $\mathbb{E}\,\|\nabla f_i(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2$ that SGD has. A simple idea that is easy to explain to undergraduates, easy to implement, and most importantly that provably brings us a better convergence rate than pure SGD and GD in a wide range of scenarios. Classical works include (Johnson and Zhang, 2013; Defazio et al., 2014). For a more thorough list of references, see the recent review (Gower et al., 2020).

**Variance reduction and VIs.** One does not need to be meticulous to quickly find finite sum problems where existing variance reduction methods do not work. In the convex world, the first that comes to mind is non-smoothness. As already mentioned, saddle point reformulations often come to rescue.

1. In the unconstrained setting, this method is also known as Optimistic Mirror Descent (OMD) or Optimistic Gradient Descent Ascent (OGDA) (Rakhlin and Sridharan, 2013; Daskalakis et al., 2018) and is also equivalent to the classical Popov's method (Popov, 1980)

| | Assumptions | Complexity |
|---|---|---|
| EG/MP, FBF, FoRB$^\dagger$ | $F$ is monotone | $\mathcal{O}\left(\frac{NL_F}{\varepsilon}\right)$ |
| EG/MP$^\ddagger$ | $F$ is monotone & $\mathbf{z} \mapsto \langle F(\mathbf{z}) + \tilde{\nabla} g(\mathbf{z}), \mathbf{z} - \mathbf{u}\rangle$ is convex for any $\mathbf{u}$ | $\mathcal{O}\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)$ |
| EG/MP$^\ddagger$ | $F$ is monotone & bounded domains | $\tilde{\mathcal{O}}\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)$ |
| FoRB$^*$ | $F$ is monotone | $\mathcal{O}\left(N + \frac{NL}{\varepsilon}\right)$ |
| **This paper** EG/MP, FBF, FoRB | $F$ is monotone | $\mathcal{O}\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)$ |

Table 1: Table of algorithms with $F(\mathbf{z}) = \sum_{i=1}^N F_i(\mathbf{z})$. EG: Extragradient, MP: Mirror-Prox, FBF: forward-backward-forward, FoRB: forward-reflected-backward. $\tilde{\nabla} g$ denotes a subgradient of $g$. $^\dagger$(Korpelevich, 1976; Tseng, 2000; Nemirovski, 2004; Malitsky and Tam, 2020), $^\ddagger$(Carmon et al., 2019), $^*$(Alacaoglu et al., 2021).

The work (Balamurugan and Bach, 2016) was seminal in using variance reduction for saddle point problems and monotone inclusions in general. In particular, the authors studied stochastic variance reduced variants of forward-backward algorithm and proved linear convergence under strong monotonicity. For bilinearly coupled problems, the complexity in (Balamurugan and Bach, 2016) improves the deterministic method in the strongly monotone setting. Chavdarova et al. (2019) developed an extragradient method with variance reduction and analyzed its convergence under strong monotonicity assumption. Unfortunately, the worst-case complexity in this work was less favorable than (Balamurugan and Bach, 2016).

Strong monotonicity may seem like a fine assumption, similar to strong convexity in minimization. While algorithmically it is true, in applications with min-max, the former is far less frequent. For instance, the operator $F$ associated with a convex-concave saddle point problem is monotone, but not strongly monotone without further assumptions. Thus, it is crucial to remove this assumption.

An influential work in this direction is by Carmon et al. (2019), where the authors proposed a randomized variant of Mirror-Prox. The authors focused primarily on matrix games and for this important case, they improved complexity over deterministic methods. However, because of this specialization, more general cases required additional assumptions. In particular, for problems beyond matrix games, the authors assumed that either $\mathbf{z} \mapsto \langle F(\mathbf{z}) + \tilde{\nabla} g(\mathbf{z}), \mathbf{z} - \mathbf{u}\rangle$ is convex for all $\mathbf{u}$ (Carmon et al., 2019, Corollary 1) or that domain is bounded (Carmon et al., 2019, Algorithm 5, Corollary 2): in particular, domain diameter is used as a parameter for this algorithm. As one can check, the former might not hold even for convex minimization problems with $F = \nabla f$. The latter, on the other hand, while already restrictive, requires a more complicated three-loop algorithm, which incurred an additional logarithmic factor into total complexity.

There are other works that did not improve complexity but introduced new ideas. An algorithm similar in spirit to ours is due to Alacaoglu et al. (2021), where variance reduction is applied to FoRB. This algorithm was the first to converge under only monotonicity, but did not improve complexity of deterministic methods. Several works studied VI methods in the stochastic setting and showed slower rates with decreasing step sizes (Mishchenko et al., 2020; Böhm et al., 2020), or

increasing mini-batch sizes (Iusem et al., 2017; Boţ et al., 2021; Cui and Shanbhag, 2021), or extra assumptions (Gorbunov et al., 2022).

### 1.2. Outline of results and comparisons

Throughout the paper, we assume access to a stochastic oracle $F_\xi$ such that $\mathbb{E}[F_\xi(\mathbf{z})] = F(\mathbf{z})$.

**Complexity and $\varepsilon$-accurate solution.** A point $\bar{\mathbf{z}}$ is an $\varepsilon$-accurate solution if $\mathbb{E}\left[\mathrm{Gap}(\bar{\mathbf{z}})\right] \leq \varepsilon$, where the gap function is defined in Section 2.3.1. Complexity of the algorithm is defined as the number of calls to $F_\xi$ to reach an $\varepsilon$-accurate solution. In general, we suppose that evaluation of $F$ is $N$ times more expensive than $F_\xi$. For specific problems with bilinear coupling, we measure the complexity in terms of arithmetic operations.

**Nonbilinear finite-sum problems.** We consider the problem (1) with $F = \sum_{i=1}^N F_i$ where $F$ is monotone, $L_F$-Lipschitz, and it is $L$-Lipschitz in mean, in view of Assumption 1(iv). In this setting, our variance reduced variants of EG, FBF, and FoRB (Corollary 6, Corollary 20, Corollary 25) have complexity $\mathcal{O}\big(N + \sqrt{N}L\varepsilon^{-1}\big)$ compared to the deterministic methods with $\mathcal{O}\left(NL_F\varepsilon^{-1}\right)$.

Our methods improve over deterministic variants as long as $L \leq \sqrt{N}L_F$. This is a similar improvement over deterministic complexity, as accelerated variance reduction does for minimization problems (Woodworth and Srebro, 2016; Allen-Zhu, 2017).

To our knowledge, the only precedent with a result similar to ours is the work (Carmon et al., 2019), where spurious assumptions were required (see Section 1.1 and Table 1), complexity had additional logarithmic terms and a complicated three-loop algorithm was needed.

**Bilinear problems.** When we focus on bilinear problems (App. D.1), the complexity of our methods is $\tilde{\mathcal{O}}\left(\mathrm{nnz}(A) + \sqrt{\mathrm{nnz}(A)(m+n)}L\varepsilon^{-1}\right)$, where $L = \|A\|_{\mathrm{Frob}}$ with Euclidean setup and $L = \|A\|_{\max}$ with simplex constraints and the entropic setup. In contrast, the complexity of deterministic method is $\tilde{\mathcal{O}}\left(\mathrm{nnz}(A)L_F\varepsilon^{-1}\right)$, where $L_F = \|A\|$ with Euclidean setup and $L_F = \|A\|_{\max}$ with the entropic setup. Our complexity shows strict improvements over deterministic methods when $A$ is dense. Our variance reduced variants for FBF and FoRB enjoy similar guarantees and obtain the same complexities (Corollary 20, Corollary 25).

In both settings this complexity was first obtained by (Carmon et al., 2019). Our results generalize the set of problems where this complexity applies due to less assumptions (for example, linearly constrained convex optimization) and also use more practical/simpler algorithms (see App. E for an empirical comparison). We also remark that our variance reduced Mirror-Prox (see Alg. 2) is different from the Mirror-Prox variant in (Carmon et al., 2019, Alg. 1, Alg. 2).

### 1.3. Organization

Most of the main body of the paper is devoted to proving the result in the case of Euclidean setup, see Section 2. These proofs contain the essential ideas that make the other results in the paper possible. In Section 3, we present our algorithm in the more general Bregman setup and highlight the main changes. The detailed proofs in this case are given in App. B. Section 4 includes an application of our results for linearly constrained convex optimization, not completely covered by previous results.

App. C includes extensions of our results to different algorithms, FBF (Tseng, 2000) and FoRB (Malitsky and Tam, 2020) which improve extragradient in terms of proximal operator's evaluation in every iteration. App. C.3 shows a linear convergence result when $g$ is strongly convex, in the

Euclidean setup. Unlike the existing results in this case, we do not require the knowledge of strong convexity parameter.

## 2. Euclidean setup

To illustrate our technique, we pick extragradient method due to the simplicity of its analysis, its extension to Bregman distances and its wide use in the literature.

### 2.1. Preliminaries

Let $\mathcal{Z}$ be a finite dimensional vector space with Euclidean inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. The notation $[N]$ represents the set $\{1, \dots, N\}$. We say $F$ is monotone if for all $\mathbf{x}, \mathbf{y}$, $\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$. Proximal operator is defined as $\operatorname{prox}_g(\mathbf{x}) = \operatorname{argmin}_{\mathbf{y}} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 \right\}$. For a proper convex lower semicontinuous (lsc) $g$, domain is defined as $\operatorname{dom} g = \{\mathbf{z} \colon g(\mathbf{z}) < +\infty\}$ and the following prox-inequality is standard

$$\bar{\mathbf{z}} = \operatorname{prox}_g(\mathbf{z}) \quad \Longleftrightarrow \quad \langle \bar{\mathbf{z}} - \mathbf{z}, \mathbf{u} - \bar{\mathbf{z}} \rangle \geq g(\bar{\mathbf{z}}) - g(\mathbf{u}), \qquad \forall \mathbf{u} \in \mathcal{Z}. \tag{3}$$

We continue with our assumptions and refer to Facchinei and Pang (2007) for sufficient conditions for Assumption 1(i).

**Assumption 1**

*(i) The solution set* Sol *of* (1) *is nonempty.*

*(ii) The function* $g \colon \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ *is proper convex lower semicontinuous.*

*(iii) The operator* $F \colon \operatorname{dom} g \to \mathcal{Z}$ *is monotone.*

*(iv) The operator* $F$ *has a stochastic oracle* $F_\xi$ *that is unbiased* $F(\mathbf{z}) = \mathbb{E}\left[F_\xi(\mathbf{z})\right]$ *and* $L$*-Lipschitz in mean:*

$$\mathbb{E}\left[\|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|^2\right] \leq L^2 \|\mathbf{u} - \mathbf{v}\|^2, \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}.$$

**Finite sum.** Suppose $F$ has a finite sum representation $F = \sum_{i=1}^N F_i$, where each $F_i$ is $L_i$-Lipschitz and the full operator $F$ is $L_F$-Lipschitz. By triangle inequality it follows, of course, that $L_F \leq \sum_{i=1}^N L_i$. On one hand, $\sum_{i=1}^N L_i$ can be much larger than $L_F$. On the other, it might be the case that $L_i$ are easy to compute, but not a true $L_F$. Then the latter inequality gives us the most natural upper bound on $L_F$. The two simplest stochastic oracles can be defined as follows

1. Uniform sampling: $F_\xi(\mathbf{z}) = N F_i(\mathbf{z})$, $q_i = \Pr\{\xi = i\} = \frac{1}{N}$. In this case, $L = \sqrt{N \sum_{i \in [N]} L_i^2}$.
2. Importance sampling: $F_\xi(\mathbf{z}) = \frac{1}{q_i} F_i(\mathbf{z})$, $q_i = \Pr\{\xi = i\} = \frac{L_i}{\sum_{j \in [N]} L_j}$. In this case, $L = \sum_{i \in [N]} L_i$.

This example is useful in several regards. First, it is one of the most general problems that proposed algorithms can tackle and for concreteness it is useful to keep it as a reference point. Second, this problem even in its generality already indicates possible pitfalls caused by non-optimal

**Algorithm 1** Extragradient with variance reduction

**Input:** Set $p \in (0,1]$, probability distribution $Q$, step size $\tau$, $\alpha \in (0,1)$, $\mathbf{z}_0 = \mathbf{w}_0$

**for** $k = 0, 1, \ldots$ **do**
$\bar{\mathbf{z}}_k = \alpha \mathbf{z}_k + (1-\alpha)\mathbf{w}_k$
$\mathbf{z}_{k+1/2} = \operatorname{prox}_{\tau g}(\bar{\mathbf{z}}_k - \tau F(\mathbf{w}_k))$
Draw an index $\xi_k$ according to $Q$
$\mathbf{z}_{k+1} = \operatorname{prox}_{\tau g}(\bar{\mathbf{z}}_k - \tau[F(\mathbf{w}_k) + F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k)])$
$\mathbf{w}_{k+1} = \begin{cases} \mathbf{z}_{k+1}, & \text{with probability } p \\ \mathbf{w}_k, & \text{with probability } 1-p \end{cases}$
**end for**

stochastic oracles. If $L$ of our stochastic oracle is much worse (meaning larger) than $L_F$, it may eliminate all advantages of cheap stochastic oracles. In the sequel, for finite-sum problems, we assume that $\xi \in [N]$, similar to the two oracles described above.

### 2.2. Extragradient with variance reduction

The classical stochastic variance reduced gradient (SVRG) (Johnson and Zhang, 2013) uses a double loop structure (looped): the full gradients are computed in the outer loop and the cheap variance reduced gradients (2) are used in the inner loop. Works (Kovalev et al., 2020; Hofmann et al., 2015) proposed a *loopless* variant of SVRG, where the outer loop was eliminated and instead full gradients were computed *once in a while* according to a randomized rule. Both methods share similar guarantees, but the latter variant is slightly simpler to analyze and implement.

We present the loopless version of extragradient with variance reduction in Alg. 1. Every iteration requires two stochastic oracles $F_\xi$ and one $F$ with probability $p$. Parameter $\alpha$ is the key in establishing a favorable complexity. While convergence of $(\mathbf{z}_k)$ to a solution will be proven for any $\alpha \in [0,1)$, a good total complexity requires a specific choice of $\alpha$. Therefore, the specific form of $\bar{\mathbf{z}}_k$ is important. Later, we see that with $\alpha = 1-p$, Alg. 1 has the claimed complexity in Table 1. It is interesting to note that by eliminating all randomness, Alg. 1 reduces to extragradient.

### 2.3. Analysis

In Alg. 1, we have two sources of randomness at each iteration: the index $\xi_k$ which is used for computing $\mathbf{z}_{k+1}$ and the choice of $\mathbf{w}_k$ (the snapshot point). We use the following notation for the conditional expectations: $\mathbb{E}[\cdot|\sigma(\xi_0, \ldots, \xi_{k-1}, \mathbf{w}_k)] = \mathbb{E}_k[\cdot]$ and $\mathbb{E}[\cdot|\sigma(\xi_0, \ldots, \xi_k, \mathbf{w}_k)] = \mathbb{E}_{k+1/2}[\cdot]$.

For the iterates $(\mathbf{z}_k)$, $(\mathbf{w}_k)$ of Alg. 1 and any $\mathbf{z} \in \operatorname{dom} g$, we define

$$\Phi_k(\mathbf{z}) := \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k - \mathbf{z}\|^2.$$

We see in the following lemma how $\Phi_k$ naturally arises in our analysis as the Lyapunov function.

**Lemma 1** *Let Assumption 1 hold, $\alpha \in [0,1)$, $p \in (0,1]$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$, for $\gamma \in (0,1)$. Then for $(\mathbf{z}_k)$ generated by Alg. 1 and any $\mathbf{z}_* \in \mathsf{Sol}$, it holds that*

$$\mathbb{E}_k\left[\Phi_{k+1}(\mathbf{z}_*)\right] \le \Phi_k(\mathbf{z}_*) - (1-\gamma)\Big((1-\alpha)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2\Big).$$

*Moreover, it holds that* $\sum_{k=0}^{\infty}\left((1-\alpha)\mathbb{E}\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right)\leq\frac{1}{1-\gamma}\Phi_0(\mathbf{z}_*)$.

**Proof** A reader may find it simpler to follow the analysis by assuming that $g$ is the indicator function of some convex set. Then since all iterates are feasible, we would have $g(\mathbf{z}_k)=0$.

Let us denote $\widehat{F}(\mathbf{z}_{k+1/2})=F(\mathbf{w}_k)+F_{\xi_k}(\mathbf{z}_{k+1/2})-F_{\xi_k}(\mathbf{w}_k)$. By prox-inequality (3) applied to the definitions of $\mathbf{z}_{k+1}$ and $\mathbf{z}_{k+1/2}$, we have that for all $\mathbf{z}$,

$$\begin{aligned}\langle\mathbf{z}_{k+1}-\bar{\mathbf{z}}_k+\tau\widehat{F}(\mathbf{z}_{k+1/2}),\mathbf{z}-\mathbf{z}_{k+1}\rangle&\geq\tau g(\mathbf{z}_{k+1})-\tau g(\mathbf{z}),\\ \langle\mathbf{z}_{k+1/2}-\bar{\mathbf{z}}_k+\tau F(\mathbf{w}_k),\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\rangle&\geq\tau g(\mathbf{z}_{k+1/2})-\tau g(\mathbf{z}_{k+1}).\end{aligned}\tag{4}$$

We sum two inequalities, use the definition of $\widehat{F}(\mathbf{z}_{k+1/2})$, and arrange to get

$$\begin{aligned}\langle\mathbf{z}_{k+1}-\bar{\mathbf{z}}_k,\mathbf{z}-\mathbf{z}_{k+1}\rangle+\langle\mathbf{z}_{k+1/2}-\bar{\mathbf{z}}_k,\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\rangle&\\ +\tau\langle F_{\xi_k}(\mathbf{w}_k)-F_{\xi_k}(\mathbf{z}_{k+1/2}),\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\rangle&\\ +\tau\langle\widehat{F}(\mathbf{z}_{k+1/2}),\mathbf{z}-\mathbf{z}_{k+1/2}\rangle\geq\tau[g(\mathbf{z}_{k+1/2})-g(\mathbf{z})].&\end{aligned}\tag{5}$$

For the first inner product we use definition of $\bar{\mathbf{z}}_k$ and identity $2\langle\mathbf{a},\mathbf{b}\rangle=\|\mathbf{a}+\mathbf{b}\|^2-\|\mathbf{a}\|^2-\|\mathbf{b}\|^2$

$$\begin{aligned}&2\langle\mathbf{z}_{k+1}-\bar{\mathbf{z}}_k,\mathbf{z}-\mathbf{z}_{k+1}\rangle=2\alpha\langle\mathbf{z}_{k+1}-\mathbf{z}_k,\mathbf{z}-\mathbf{z}_{k+1}\rangle+2(1-\alpha)\langle\mathbf{z}_{k+1}-\mathbf{w}_k,\mathbf{z}-\mathbf{z}_{k+1}\rangle\\ &=\alpha\left(\|\mathbf{z}_k-\mathbf{z}\|^2-\|\mathbf{z}_{k+1}-\mathbf{z}\|^2-\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2\right)\\ &\quad+(1-\alpha)\left(\|\mathbf{w}_k-\mathbf{z}\|^2-\|\mathbf{z}_{k+1}-\mathbf{z}\|^2-\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2\right)\\ &=\alpha\|\mathbf{z}_k-\mathbf{z}\|^2-\|\mathbf{z}_{k+1}-\mathbf{z}\|^2+(1-\alpha)\|\mathbf{w}_k-\mathbf{z}\|^2-\alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2-(1-\alpha)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2.\end{aligned}\tag{6}$$

Similarly, for the second inner product in (5) we deduce

$$\begin{aligned}2\langle\mathbf{z}_{k+1/2}-\bar{\mathbf{z}}_k,\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\rangle=\alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2-\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2+(1-\alpha)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2&\\ -\alpha\|\mathbf{z}_{k+1/2}-\mathbf{z}_k\|^2-(1-\alpha)\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2.&\end{aligned}\tag{7}$$

For the remaining terms in (5), we plug in $\mathbf{z}=\mathbf{z}_*$, use that $\mathbf{z}_{k+1/2},\mathbf{w}_k$ is deterministic under the conditioning of $\mathbb{E}_k$ and $\mathbb{E}_k[\widehat{F}(\mathbf{z}_{k+1/2})]=\mathbb{E}_k\left[F(\mathbf{w}_k)+F_{\xi_k}(\mathbf{z}_{k+1/2})-F_{\xi_k}(\mathbf{w}_k)\right]=F(\mathbf{z}_{k+1/2})$ to obtain

$$\begin{aligned}&\mathbb{E}_k\left[\langle\widehat{F}(\mathbf{z}_{k+1/2}),\mathbf{z}_*-\mathbf{z}_{k+1/2}\rangle+g(\mathbf{z}_*)-g(\mathbf{z}_{k+1/2})\right]\\ &=\langle F(\mathbf{z}_{k+1/2}),\mathbf{z}_*-\mathbf{z}_{k+1/2}\rangle+g(\mathbf{z}_*)-g(\mathbf{z}_{k+1/2})&&(\mathbb{E}_k[F(\mathbf{w}_k)-F_{\xi_k}(\mathbf{w}_k)]=0)\\ &\leq\langle F(\mathbf{z}_*),\mathbf{z}_*-\mathbf{z}_{k+1/2}\rangle+g(\mathbf{z}_*)-g(\mathbf{z}_{k+1/2})\leq 0&&(\text{monotonicity and (1)})\end{aligned}\tag{8}$$

and

$$\begin{aligned}&\mathbb{E}_k\left[2\tau\langle F_{\xi_k}(\mathbf{w}_k)-F_{\xi_k}(\mathbf{z}_{k+1/2}),\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\rangle\right]\\ &\leq\mathbb{E}_k\left[2\tau\|F_{\xi_k}(\mathbf{w}_k)-F_{\xi_k}(\mathbf{z}_{k+1/2})\|\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|\right]&&(\text{Cauchy-Schwarz})\\ &\leq\frac{\tau^2}{\gamma}\mathbb{E}_k\left[\|F_{\xi_k}(\mathbf{z}_{k+1/2})-F_{\xi_k}(\mathbf{w}_k)\|^2\right]+\gamma\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right]&&(\text{Young's ineq.})\\ &\leq(1-\alpha)\gamma\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\gamma\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right].&&(\text{Assumption 1(iv)})\end{aligned}\tag{9}$$

We use (6), (7), (8), and (9) in (5), after taking expectation $\mathbb{E}_k$ and letting $\mathbf{z} = \mathbf{z}_*$, to deduce

$$\mathbb{E}_k\left[\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2\right] \le \alpha\|\mathbf{z}_k - \mathbf{z}_*\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}_*\|^2 - (1-\alpha)(1-\gamma)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 - (1-\gamma)\,\mathbb{E}_k\left[\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2\right]. \tag{10}$$

By the definition of $\mathbf{w}_{k+1}$ and $\mathbb{E}_{k+1/2}$, it follows that

$$\frac{1-\alpha}{p}\mathbb{E}_{k+1/2}\left[\|\mathbf{w}_{k+1} - \mathbf{z}_*\|^2\right] = (1-\alpha)\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2 + (1-\alpha)\left(\frac{1}{p} - 1\right)\|\mathbf{w}_k - \mathbf{z}_*\|^2. \tag{11}$$

We add (11) to (10) and apply the tower property $\mathbb{E}_k\left[\mathbb{E}_{k+1/2}[\cdot]\right] = \mathbb{E}_k[\cdot]$ to deduce

$$\alpha\mathbb{E}_k\left[\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2\right] + \frac{1-\alpha}{p}\mathbb{E}_k\left[\|\mathbf{w}_{k+1} - \mathbf{z}_*\|^2\right] \le \alpha\|\mathbf{z}_k - \mathbf{z}_*\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k - \mathbf{z}_*\|^2 - (1-\gamma)\Big((1-\alpha)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 + \mathbb{E}_k\left[\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2\right]\Big).$$

Using the definition of $\Phi_k(\mathbf{z})$, we obtain the first result. Applying total expectation and summing the inequality yields the second result. ■

To show the almost sure convergence of the sequence $(\mathbf{z}_k)$, we need $F_\xi$ to be continuous for all $\xi$. For a finite sum example it follows automatically from Assumption 1. The proof is given in App. A.2.

**Theorem 2** *Let Assumption 1 hold, $F_\xi$ be continuous for all $\xi$, $\alpha \in [0,1)$, $p \in (0,1]$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$, for $\gamma \in (0,1)$. Then, almost surely there exists $\mathbf{z}_* \in \mathsf{Sol}$ such that $(\mathbf{z}_k)$ generated by Alg. 1 converges to $\mathbf{z}_*$.*

### 2.3.1. CONVERGENCE RATE AND COMPLEXITY FOR MONOTONE CASE

In the general monotone case, the convergence measure is the gap function given by

$$\mathrm{Gap}(\mathbf{w}) = \max_{\mathbf{z}\in\mathcal{C}}\big\{\langle F(\mathbf{z}), \mathbf{w} - \mathbf{z}\rangle + g(\mathbf{w}) - g(\mathbf{z})\big\},$$

where $\mathcal{C}$ is a compact subset of $\mathcal{Z}$ that we use to handle the possibility of unboundedness of $\operatorname{dom} g$ (see (Nesterov, 2007, Lemma 1)). Since we work in probabilistic setting, naturally our convergence measure will be based on $\mathbb{E}[\mathrm{Gap}(\mathbf{w})]$. We start with a simple lemma for "switching" the order of maximum and expectation, which is required for showing convergence of expected gap. This technique is standard for such purpose Nemirovski et al. (2009) and the proof is given in App. A.1.

**Lemma 3** *Let $\mathcal{F} = (\mathcal{F}_k)_{k\ge 0}$ be a filtration and $(\mathbf{u}_k)$ a stochastic process adapted to $\mathcal{F}$ with $\mathbb{E}[\mathbf{u}_{k+1}|\mathcal{F}_k] = 0$. Then for any $K \in \mathbb{N}$, $\mathbf{x}_0 \in \mathcal{Z}$, and any compact set $\mathcal{C} \subset \mathcal{Z}$,*

$$\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{C}}\sum_{k=0}^{K-1}\langle \mathbf{u}_{k+1}, \mathbf{x}\rangle\right] \le \max_{\mathbf{x}\in\mathcal{C}}\frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}\|^2 + \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{u}_{k+1}\|^2.$$

We now continue with the main result of this section.

**Theorem 4** *Let Assumption 1 hold, $p \in (0,1]$, $\alpha = 1-p$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$, for $\gamma \in (0,1)$. Then, for $\mathbf{z}^K = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{z}_{k+1/2}$, it follows that*

$$\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] = \mathcal{O}\left(\frac{L}{\sqrt{p}K}\right).$$

*In particular, for $\tau = \frac{\sqrt{p}}{2L}$, the rate is $\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \leq \frac{17.5L}{\sqrt{p}K}\max_{z\in\mathcal{C}}\|\mathbf{z}_0 - \mathbf{z}\|^2$.*

Recall that we measure complexity in terms of calls to the stochastic oracle $F_\xi(\cdot)$ and we assumed that the cost of computing $F(\cdot)$ is $N$ times that of $F_\xi(\cdot)$. For a finite sum example, this is a natural assumption. Below we provide a proof sketch of the theorem and the full proof is given in App. A.1, along with the proof of Lemma 3.

**Remark 5** *For Alg. 1, since per iteration cost is $pN+2$ calls to $F_\xi$ in expectation, the result is "average" total complexity: expected number of calls to get a small expected gap.*

**Corollary 6** *In the setting of Theorem 4, the average total complexity of Alg. 1 to reach $\varepsilon$-accuracy is $\mathcal{O}\left(N + (pN+2)\left(1 + \frac{L}{\sqrt{p}\varepsilon}\right)\right)$. In particular, for $p = \frac{2}{N}$ it is $\mathcal{O}\left(N + \frac{\sqrt{N}L}{\varepsilon}\right)$.*

**Proof sketch of Theorem 4** As already mentioned, when all randomness is eliminated, that is $F_\xi = F$ and $p = 1$, Alg. 1 reduces to extragradient. In that case, the convergence rate $\mathcal{O}(1/K)$ would follow almost immediately from the proof of Lemma 1. In the stochastic setting the proof is more subtle and we have to rely on Lemma 3 to deal with the error terms caused by randomness. Let

$$\Theta_{k+1/2}(\mathbf{z}) = \langle F(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1/2} - \mathbf{z}\rangle + g(\mathbf{z}_{k+1/2}) - g(\mathbf{z}).$$

We will proceed as in Lemma 1 before getting (10). In particular, using (6) and (7) in (5), using the definition of $\Theta_{k+1/2}$ and $\Phi_k(\mathbf{z}) = (1-p)\|\mathbf{z}_k - \mathbf{z}\|^2 + \|\mathbf{w}_k - \mathbf{z}\|^2$ with $\alpha = 1-p$ gives

$$\begin{aligned}2\tau\Theta_{k+1/2}(\mathbf{z}) + \Phi_{k+1}(\mathbf{z}) \leq \Phi_k(\mathbf{z}) &+ e_1(\mathbf{z},k) + e_2(\mathbf{z},k)\\ &+ 2\tau\langle F_{\xi_k}(\mathbf{w}_k) - F_{\xi_k}(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\rangle\\ &- p\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2,\end{aligned} \tag{12}$$

where we defined the error terms

$$\begin{aligned}e_1(\mathbf{z},k) &= 2\tau\langle F(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{z}_{k+1/2}) - F(\mathbf{w}_k) + F_{\xi_k}(\mathbf{w}_k), \mathbf{z}_{k+1/2} - \mathbf{z}\rangle,\\ e_2(\mathbf{z},k) &= p\|\mathbf{w}_k - \mathbf{z}\|^2 + \|\mathbf{w}_{k+1} - \mathbf{z}\|^2 - \|\mathbf{w}_k - \mathbf{z}\|^2 - p\|\mathbf{z}_{k+1} - \mathbf{z}\|^2\\ &= 2\langle p\mathbf{z}_{k+1} + (1-p)\mathbf{w}_k - \mathbf{w}_{k+1}, \mathbf{z}\rangle - p\|\mathbf{z}_{k+1}\|^2 - (1-p)\|\mathbf{w}_k\|^2 + \|\mathbf{w}_{k+1}\|^2.\end{aligned} \tag{13}$$

We sum (12) over $k = 0, \ldots, K-1$, take maximum over $\mathbf{z} \in \mathcal{C}$, and take total expectation to get

$$\begin{aligned}2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \leq \max_{\mathbf{z}\in\mathcal{C}}\Phi_0(\mathbf{z}) &+ \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\left(e_1(\mathbf{z},k) + e_2(\mathbf{z},k)\right)\right]\\ &- \mathbb{E}\sum_{k=0}^{K-1}\left(\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 + p\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2\right)\\ &+ 2\tau\mathbb{E}\sum_{k=0}^{K-1}\left[\langle F_{\xi_k}(\mathbf{w}_k) - F_{\xi_k}(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\rangle\right]\end{aligned} \tag{14}$$

where we used $\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \Theta_{k+1/2}(\mathbf{z})\right] \geq K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right]$, which follows from monotonicity of $F$, linearity of $\langle F(\mathbf{z}), \cdot - \mathbf{z}\rangle$ for any $\mathbf{z}$, and convexity of $g$.

The tower property, the estimation from (9), and $1-\alpha = p$ applied on (14) imply

$$2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \leq \max_{\mathbf{z}\in\mathcal{C}} \Phi_0(\mathbf{z}) + \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \big(e_1(\mathbf{z},k) + e_2(\mathbf{z},k)\big)\right]. \tag{15}$$

Therefore, the proof will be complete upon deriving an upper bound for the second term on RHS. We instantiate Lemma 3 twice for bounding this term. First, for $e_1(\mathbf{z},k)$, Lemma 3 implies

$$\mathbb{E}\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} e_1(\mathbf{z},k) \leq \max_{z\in\mathcal{C}} \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 + 2\tau^2 L^2 \sum_{k=0}^{K-1} \mathbb{E}\,\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2. \tag{16}$$

Secondly, for $e_2(\mathbf{z},k)$, Lemma 3 implies

$$\mathbb{E}\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} e_2(\mathbf{z},k) \leq \max_{\mathbf{z}\in\mathcal{C}} \|\mathbf{z}_0 - \mathbf{z}\|^2 + p(1-p) \sum_{k=0}^{K-1} \mathbb{E}\,\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2. \tag{17}$$

We combine (16), (17), and (15), and use the second result of Lemma 1, to estimate terms

$$\mathbb{E}\left[\sum_{k=0}^{K-1} \left(2\tau^2 L^2 \|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 + p(1-p)\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2\right)\right] \leq \frac{3.5}{1-\gamma} \max_{\mathbf{z}\in\mathcal{C}} \Phi_0(\mathbf{z}). \tag{18}$$

By using $\mathbf{w}_0 = \mathbf{z}_0$ and $\tau = \frac{\sqrt{p}\gamma}{L}$ and straightforward calculations, we finish the proof. ■

## 3. Bregman setup

### 3.1. Preliminaries

In this section, we assume that $\mathcal{Z}$ is a normed vector space with a dual space $\mathcal{Z}^*$ and primal-dual norm pair $\|\cdot\|$ and $\|\cdot\|_*$. Let $h\colon \mathcal{Z} \to \mathbb{R}\cup\{+\infty\}$ be a proper convex lsc function that satisfies *(i)* $\operatorname{dom} g \subseteq \operatorname{dom} h$, *(ii)* $h$ is differentiable over $\operatorname{dom}\partial h$, *(iii)* $h$ is 1-strongly convex on $\operatorname{dom} g$. Then we can define the Bregman distance $D\colon \operatorname{dom} g \times \operatorname{dom}\partial h \to \mathbb{R}_+$ associated with $h$ by

$$D(\mathbf{u},\mathbf{v}) := h(\mathbf{u}) - h(\mathbf{v}) - \langle \nabla h(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle.$$

Note that since $h$ is 1-strongly convex with respect to norm $\|\cdot\|$, we have $D(\mathbf{u},\mathbf{v}) \geq \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|^2$.

Naturally, we shall say that $F\colon \operatorname{dom} g \to \mathcal{Z}^*$ is $L_F$-Lipschitz, if $\|F(\mathbf{u}) - F(\mathbf{v})\|_* \leq L_F\|\mathbf{u} - \mathbf{v}\|, \forall \mathbf{u},\mathbf{v}$. However, Lipschitzness for a stochastic oracle this time will be more involved. Evidently, we prefer stochastic oracles $F_\xi$ of $F$ with as small $L$ as possible. Moreover, the proof of Lemma 1 indicates that in $k$-th iteration we need Lipschitzness only for already known two iterates. Hence, following (Grigoriadis and Khachiyan, 1995; Carmon et al., 2019), in contrast to Alg. 1, we will not fix distribution $Q$ in the beginning, but allow it to vary from iteration to iteration. Formally, this amounts to the following definition.

**Algorithm 2** Mirror-prox with variance reduction

**Input:** Step size $\tau$, $\alpha \in (0,1)$, $K > 0$. Let $\mathbf{z}_j^{-1} = \mathbf{z}_0^0 = \mathbf{w}^0 = \mathbf{z}_0, \forall j \in [K]$
**for** $s = 0, 1 \dots$ **do**
  **for** $k = 0, 1 \dots K-1$ **do**
    $\mathbf{z}_{k+1/2}^s = \operatorname{argmin}_{\mathbf{z}} \Big\{ g(\mathbf{z}) + \langle F(\mathbf{w}^s), \mathbf{z} \rangle + \frac{\alpha}{\tau} D(\mathbf{z}, \mathbf{z}_k^s) + \frac{1-\alpha}{\tau} D(\mathbf{z}, \bar{\mathbf{w}}^s) \Big\}$.
    Fix distribution $Q_{\mathbf{z}_{k+1/2}^s, \mathbf{w}^s}$ and sample $\xi_k^s$ according to it
    $\widehat{F}(\mathbf{z}_{k+1/2}^s) = F(\mathbf{w}^s) + F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s)$
    $\mathbf{z}_{k+1}^s = \operatorname{argmin}_{\mathbf{z}} \Big\{ g(\mathbf{z}) + \langle \widehat{F}(\mathbf{z}_{k+1/2}^s), \mathbf{z} \rangle + \frac{\alpha}{\tau} D(\mathbf{z}, \mathbf{z}_k^s) + \frac{1-\alpha}{\tau} D(\mathbf{z}, \bar{\mathbf{w}}^s) \Big\}$.
  **end for**
  $\mathbf{w}^{s+1} = \frac{1}{K} \sum_{k=1}^K \mathbf{z}_k^s$
  $\nabla h(\bar{\mathbf{w}}^{s+1}) = \frac{1}{K} \sum_{k=1}^K \nabla h(\mathbf{z}_k^s)$
  $\mathbf{z}_0^{s+1} = \mathbf{z}_K^s$
**end for**

**Definition 7** *We say that $F$ has a stochastic oracle $F_\xi$ that is* variable *$L$-Lipschitz in mean, if for any $\mathbf{u}, \mathbf{v} \in \operatorname{dom} g$ there exists a distribution $Q_{\mathbf{u},\mathbf{v}}$ such that*

*(i) $F$ is unbiased:* $F(\mathbf{z}) = \mathbb{E}_{\xi \sim Q_{\mathbf{u},\mathbf{v}}} [F_\xi(\mathbf{z})] \quad \forall \mathbf{z} \in \operatorname{dom} g$;

*(ii)* $\mathbb{E}_{\xi \sim Q_{\mathbf{u},\mathbf{v}}} \left[ \|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|_*^2 \right] \le L^2 \|\mathbf{u} - \mathbf{v}\|^2$.

Note that the second condition holds only for given $\mathbf{u}, \mathbf{v}$, but the constant $L$ is universal for all $\mathbf{u}, \mathbf{v}$. Changing $\mathbf{u}, \mathbf{v}$ also changes a distribution, hence the name "variable". Without loss of generality, we denote any distribution that realizes the above Lipschitz bound for given $\mathbf{u}$, $\mathbf{v}$ by $Q_{\mathbf{u},\mathbf{v}}$. This definition resembles the one in (Carmon et al., 2019, Definition 2). It is easy to see when $Q_{\mathbf{u},\mathbf{v}} = Q$ for all $\mathbf{u}, \mathbf{v}$, we get the same definition as before in Assumption 1.

We now introduce Assumption 2 which will replace and generalize Assumption 1(iv).

**Assumption 2** *The operator $F$ has a stochastic oracle $F_\xi$ that is variable $L$-Lipschitz in mean (see Definition 7).*

### 3.2. Mirror-Prox with variance reduction

In this setting, we could simply adjust the steps of Alg. 1 and correspondingly the analysis of Lemma 1. However, to show a convergence rate, double randomization in Alg. 1 causes technical complications. For this reason, in the Bregman setup we propose a double loop variant of Alg. 1, similar to the classical SVRG (Johnson and Zhang, 2013). Our algorithm can be seen as a variant of Mirror-Prox (Nemirovski, 2004) with variance reduction. Now it should be clear that Alg. 1 is a randomized version of Alg. 2 with $p = \frac{1}{K}$ and a particular choice $D(\mathbf{z}, \mathbf{z}') = \frac{1}{2}\|\mathbf{z} - \mathbf{z}'\|_2^2$.

The technical reason for this change is the calculation given in (13). In fact, all the other steps in the previous proofs would go through by using three point identity (see (30)), except this step, which is inherently using the properties of $\ell_2$-norm. By removing double randomization and introducing double loop instead, step (13) will not be needed in the analysis of Bregman case.

Compared to Alg. 1, $\mathbf{w}^s$ serves the same purpose as $\mathbf{w}_k$: the snapshot point in the language of SVRG (Johnson and Zhang, 2013). Since we have two loops in this case, we get $\mathbf{w}^s$ by averaging, again, similar to SVRG for non-strongly convex optimization (Reddi et al., 2016; Allen-Zhu and Yuan, 2016). The difference due to Bregman setup is that we have the additional point $\bar{\mathbf{w}}^s$ that averages in the dual space. This operation does not incur additional cost. Proofs of the results in this section are given in App. B.

**Theorem 8** *Let Assumption 1(i,ii,iii) and Assumption 2 hold, $\alpha \in [0,1)$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ for $\gamma \in (0,1)$. Then, for $\mathbf{z}^S = \frac{1}{KS}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\mathbf{z}^s_{k+1/2}$, it follows that*

$$\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^S)\right] \le \frac{1}{\tau KS}\left(1 + \left(1 + \frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha))\right)\max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0).$$

**Corollary 9** *Let $K = \frac{N}{2}$ and $\alpha = 1 - \frac{1}{K} = 1 - \frac{2}{N}$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ for $\gamma \in (0,1)$. Then the total complexity of Alg. 2 to reach $\varepsilon$-accuracy is $\mathcal{O}\left(N + \frac{L\sqrt{N}}{\varepsilon}\right)$. In particular, if $\tau = \frac{\sqrt{1-\alpha}}{3L} = \frac{\sqrt{2}}{3\sqrt{N}L}$, the total complexity is $2N + \frac{43\sqrt{N}L}{\varepsilon}\max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0)$.*

## 4. Application: Linearly constrained minimization

A classical example of bilinear saddle point problems is linearly constrained minimization

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{subject to} \quad A\mathbf{x} = b,$$

where $f$ is proper convex lsc. The equivalent min-max formulation corresponds to (1) with $F(\mathbf{x},\mathbf{y}) = (A^\top\mathbf{y}, -A\mathbf{x})$ and $g(\mathbf{x},\mathbf{y}) = f(\mathbf{x}) + \langle b, \mathbf{y}\rangle$. For $A \in \mathbb{R}^{m\times n}$ we denote a number of its non-zero entries by $\mathrm{nnz}(A)$. The spectral and Frobenius norms of $A$ are denoted as $\|A\|$ and $\|A\|_{\mathrm{Frob}}$. For $i$-th row and $j$-th column of $A$ we use a convenient notation $A_{i:}$ and $A_{:j}$.

We instantiate Alg. 1 for this problem. To make our presentation clearer, we consider only the most common scenario when $\mathrm{nnz}(A) > m+n$. In this setting, deterministic methods (extragradient, FBF, FoRB, etc.) solve (62) with $\mathcal{O}\left(\mathrm{nnz}(A)\|A\|\varepsilon^{-1}\right)$ total complexity. As we see now, variance reduced methods provide us $\mathcal{O}\left(\mathrm{nnz}(A) + \sqrt{\mathrm{nnz}(A)(m+n)}\|A\|_{\mathrm{Frob}}\varepsilon^{-1}\right)$ total complexity.

The oracle is defined as

$$F_\xi(\mathbf{z}) = \begin{pmatrix} \frac{1}{r_i}A_{i:}y_i \\ -\frac{1}{c_j}A_{:j}x_j \end{pmatrix}, \quad \Pr\{\xi = (i,j)\} = r_i c_j, \quad r_i = \frac{\|A_{i:}\|_2^2}{\|A\|_{\mathrm{Frob}}^2}, \quad c_j = \frac{\|A_{:j}\|_2^2}{\|A\|_{\mathrm{Frob}}^2}.$$

In view of Assumption 1(iv), $F_\xi$ is $\|A\|_{\mathrm{Frob}}$-Lipschitz in mean (see (60) for the derivation). For an alternative oracle, see the extended discussion in App. D.1.1.

**Complexity.** We suppose that computing $\mathrm{prox}_f$ can be done efficiently in $\tilde{\mathcal{O}}(m+n)$ complexity. Our result in Theorem 4 stated that Alg. 1 has the rate $\mathcal{O}\left(\frac{L}{\sqrt{p}K}\right)$. Given that the expected cost of each iteration is $\mathcal{O}\left(p\,\mathrm{nnz}(A) + m + n\right)$, setting $p = \frac{m+n}{\mathrm{nnz}(A)}$ gives us the average total complexity

$$\tilde{\mathcal{O}}\left(\mathrm{nnz}(A) + \frac{\sqrt{\mathrm{nnz}(A)(m+n)}\|A\|_{\mathrm{Frob}}}{\varepsilon}\right). \tag{19}$$

It is easy to see that Alg. 2 has the same complexity if we set $K = \left\lceil \frac{\mathrm{nnz}(A)}{m+n} \right\rceil$. Compared to the complexity of deterministic methods, this complexity improves depending on the relation between $\|A\|_{\mathrm{Frob}}$ and $\|A\|$. In particular, when $A$ is a square dense matrix, due to $\|A\|_{\mathrm{Frob}} \leq \sqrt{\mathrm{rank}(A)}\|A\|$, the bound in (19) improves that of deterministic method. In (19) we suppress $\|\mathbf{z}_0 - \mathbf{z}_*\|^2$ that is common to all methods considered in this paragraph.

Finally, we remark that the analysis in (Carmon et al., 2019, Section 5.2) requires the additional assumption that $\mathbf{z} \mapsto \langle F(\mathbf{z}) + \tilde{\nabla} f(\mathbf{z}), \mathbf{z} - \mathbf{u} \rangle$ is convex for all $\mathbf{u}$ to apply to this case, where we denote a subgradient of $f$ by $\tilde{\nabla} f$. This assumption requires more structure on $f$.

## 5. Conclusions

We conclude by discussing a few potential directions that our results could pave the way for.

**Sparsity.** An important consideration in practice is to adapt to sparsity of the data. The recent work by Carmon et al. (2020) built on the algorithm in (Carmon et al., 2019) and improved the complexity for matrix games in Euclidean setup, for sparse data, by using specialized data structures. We suspect that these techniques can also be used in our algorithms.

**Stochastic oracles.** As we have seen for bilinear and nonbilinear problems, harnessing the structure is very important for devising suitable stochastic oracles with small Lipschitz constants. On top of our algorithms, an interesting direction is to study important nonbilinear min-max problems and devise particular Bregman distances and stochastic oracles to obtain complexity improvements.

**New algorithms.** For brevity, we only showed the application of our techniques for extragradient, FBF, and FoRB methods. However, for more structured problems other extensions might be more suitable. Such structured problems arise, for example, when only partial strong convexity is present or when $F$ is the sum of a skew-symmetric matrix and a gradient of a convex function.

## Acknowledgments

Part of the work was done while A. Alacaoglu was at EPFL. The work of A. Alacaoglu has received funding from the NSF Award 2023239; DOE ASCR under Subcontract 8F-30039 from Argonne National Laboratory; the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 725594 - time-data). The work of Y. Malitsky was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The project number is 305286.

## References

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

Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Z. Allen-Zhu and Y. Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International Conference on Machine Learning*, pages 1080–1089. PMLR, 2016.

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

A. Böhm, M. Sedlmayer, E. R. Csetnek, and R. I. Boţ. Two steps at a time – taking GAN training in stride with Tseng's method. *arXiv:2006.09033*, 2020.

R. I. Boţ, 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.

Y. Carmon, Y. Jin, A. Sidford, and K. Tian. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems*, pages 11377–11388, 2019.

Y. Carmon, Y. Jin, A. Sidford, and K. Tian. Coordinate methods for matrix games. In *IEEE 61st Annual Symposium on Foundations of Computer Science*, pages 283–293. IEEE, 2020.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, pages 391–401, 2019.

K. L. Clarkson, E. Hazan, and D. P. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM*, 59(5):1–49, 2012.

P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

L. Condat. Fast projection onto the simplex and the $\ell_1$ ball. *Mathematical Programming*, 158(1): 575–585, 2016.

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*, 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*, pages 1646–1654, 2014.

E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3 (4):1015–1046, 2010.

F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

I. Gemp and S. Mahadevan. Global convergence to the equilibrium of GANs using variational inequalities. *arXiv:1808.01531*, 2018.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In *International Conference on Learning Representations*, 2019.

N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.

R. M. Gower, M. Schmidt, F. Bach, and P. Richtarik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58, 1995.

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

T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekon. Mat. Metody*, 12:747–756, 1976.

D. Kovalev, S. Horvath, and P. Richtárik. Don't Jump Through Hoops and Remove Those Loops: SVRG and Katyusha are Better Without the Outer Loop. In *International Conference on Algorithmic Learning Theory*, pages 451–467, 2020.

Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

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

P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

A. Nemirovski. Mini-Course on Convex Programming Algorithms. Lecture notes, 2013.

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

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1): 127–152, 2005.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Y. Nesterov and A. Nemirovski. On first order algorithms for $\ell_1$/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2016.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. 1971.

P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

B. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pages 3646–3654, 2016.

H. Zhang. Extragradient and extrapolation methods with generalized bregman distances for saddle point problems. *Operations Research Letters*, 50(3):329–334, 2022.

## Appendix A. Appendix for Section 2

**Remark 10** *For running Alg. 1 in practice, we suggest* $p = \frac{2}{N}$, $\alpha = 1 - p$, *and* $\tau = \frac{0.99\sqrt{p}}{L}$. *Specific problem may require a more careful examination of "optimal" parameters (see App. D.1).*

**Remark 11** *Our rate guarantee on Theorem 4 is on the averaged iterate* $\mathbf{z}^K$, *which is shown to be necessary to get the* $O(1/K)$ *rate for deterministic extragradient in (Golowich et al., 2020).*

### A.1. Full proof of Theorem 4

**Proof of Lemma 3** First, we define the sequence $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{u}_{k+1}$. It is easy to see that $\mathbf{x}_k$ is $\mathcal{F}_k$-measurable. Next, by using the definition of $(\mathbf{x}_k)$, we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}\|^2 = \|\mathbf{x}_k - \mathbf{x}\|^2 + 2\langle \mathbf{u}_{k+1}, \mathbf{x}_k - \mathbf{x}\rangle + \|\mathbf{u}_{k+1}\|^2.$$

Summing over $k = 0, \ldots, K-1$, we obtain

$$\sum_{k=0}^{K-1} 2\langle \mathbf{u}_{k+1}, \mathbf{x} - \mathbf{x}_k\rangle \leq \|\mathbf{x}_0 - \mathbf{x}\|^2 + \sum_{k=0}^{K-1} \|\mathbf{u}_{k+1}\|^2.$$

Next, we take maximum of both sides and then expectation

$$\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{C}} \sum_{k=0}^{K-1} \langle \mathbf{u}_k, \mathbf{x}\rangle\right] \leq \max_{\mathbf{x}\in\mathcal{C}} \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}\|^2 + \frac{1}{2}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\mathbf{u}_{k+1}\|^2\right] + \sum_{k=0}^{K-1} \mathbb{E}\left[\langle \mathbf{u}_{k+1}, \mathbf{x}_k\rangle\right].$$

We use the tower property, $\mathcal{F}_k$-measurability of $\mathbf{x}_k$, and $\mathbb{E}\left[\mathbf{u}_{k+1}|\mathcal{F}_k\right] = 0$ to finish the proof, since $\sum_{k=0}^{K-1} \mathbb{E}\left[\langle \mathbf{u}_{k+1}, \mathbf{x}_k\rangle\right] = \sum_{k=0}^{K-1} \mathbb{E}\left[\langle \mathbb{E}\left[\mathbf{u}_{k+1}|\mathcal{F}_k\right], \mathbf{x}_k\rangle\right] = 0$. ■

**Proof of Theorem 4** As already mentioned, when all randomness is eliminated, that is $F_\xi = F$ and $p = 1$, Alg. 1 reduces to extragradient. In that case, the convergence rate $\mathcal{O}(1/K)$ would follow almost immediately from the proof of Lemma 1. In the stochastic setting the proof is more subtle and we have to rely on Lemma 3 to deal with the error terms caused by randomness. Let us recall $\widehat{F}(\mathbf{z}_{k+1/2}) = F(\mathbf{w}_k) + F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k)$ and

$$\Theta_{k+1/2}(\mathbf{z}) = \langle F(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1/2} - \mathbf{z}\rangle + g(\mathbf{z}_{k+1/2}) - g(\mathbf{z}).$$

We will proceed as in Lemma 1 before getting (10). In particular, using (6) and (7) in (5) gives

$$\begin{aligned} 2\tau\Theta_{k+1/2}(\mathbf{z}) + \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 \leq{}& \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}\|^2 \\ &+ 2\tau\langle F_{\xi_k}(\mathbf{w}_k) - F_{\xi_k}(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\rangle \\ &- (1-\alpha)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 \\ &+ \underbrace{2\tau\langle F(\mathbf{z}_{k+1/2}) - \widehat{F}(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1/2} - \mathbf{z}\rangle}_{e_1(\mathbf{z},k)}, \end{aligned} \tag{20}$$

where we call the last term by $e_1(\mathbf{z}, k)$.

Now, we set $\alpha = 1-p$. We want to rewrite (20) using $\Phi_k(\mathbf{z}) = (1-p)\|\mathbf{z}_k - \mathbf{z}\|^2 + \|\mathbf{w}_k - \mathbf{z}\|^2$. For this, we need to add $\|\mathbf{w}_{k+1} - \mathbf{z}\|^2 - \|\mathbf{w}_k - \mathbf{z}\|^2$ to both sides. Then, we define the error

$$\begin{aligned} e_2(\mathbf{z}, k) &= p\|\mathbf{w}_k - \mathbf{z}\|^2 + \|\mathbf{w}_{k+1} - \mathbf{z}\|^2 - \|\mathbf{w}_k - \mathbf{z}\|^2 - p\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 \\ &= 2\langle p\mathbf{z}_{k+1} + (1-p)\mathbf{w}_k - \mathbf{w}_{k+1}, \mathbf{z}\rangle - p\|\mathbf{z}_{k+1}\|^2 - (1-p)\|\mathbf{w}_k\|^2 + \|\mathbf{w}_{k+1}\|^2. \end{aligned}$$

With this at hand, we can cast (20) as

$$\begin{aligned} 2\tau\Theta_{k+1/2}(\mathbf{z}) + \Phi_{k+1}(\mathbf{z}) \le \Phi_k(\mathbf{z}) + e_1(\mathbf{z}, k) + e_2(\mathbf{z}, k) \\ + 2\tau\langle F_{\xi_k}(\mathbf{w}_k) - F_{\xi_k}(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\rangle \\ - p\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2. \end{aligned}$$

We sum this inequality over $k = 0, \dots, K-1$, take maximum of both sides over $\mathbf{z} \in \mathcal{C}$, and then take total expectation to obtain

$$\begin{aligned} 2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \le \max_{\mathbf{z}\in\mathcal{C}} \Phi_0(\mathbf{z}) + \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \left(e_1(\mathbf{z}, k) + e_2(\mathbf{z}, k)\right)\right] \\ - \mathbb{E}\sum_{k=0}^{K-1} \left(\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 + p\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2\right) \\ + 2\tau\mathbb{E}\sum_{k=0}^{K-1} \left[\langle F_{\xi_k}(\mathbf{w}_k) - F_{\xi_k}(\mathbf{z}_{k+1/2}), \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\rangle\right] \end{aligned} \tag{21}$$

where we used $\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \Theta_{k+1/2}(\mathbf{z})\right] \ge K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right]$, which follows from monotonicity of $F$, linearity of $\langle F(\mathbf{z}), \cdot - \mathbf{z}\rangle$ for any $\mathbf{z}$, and convexity of $g$.

The tower property, the estimation from (9), and $1 - \alpha = p$ applied on (21) imply

$$2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \le \max_{\mathbf{z}\in\mathcal{C}} \Phi_0(\mathbf{z}) + \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \left(e_1(\mathbf{z}, k) + e_2(\mathbf{z}, k)\right)\right]. \tag{22}$$

Therefore, the proof will be complete upon deriving an upper bound for the second term on RHS. We now instantiate Lemma 3 twice for bounding this term. First, for $e_1(\mathbf{z}, k)$ we set in Lemma 3,

$$\mathcal{F}_k = \sigma(\xi_0, \dots, \xi_{k-1}, \mathbf{w}_k), \quad \tilde{\mathbf{x}}_0 = \mathbf{z}_0, \quad \mathbf{u}_{k+1} = 2\tau\big(\widehat{F}(\mathbf{z}_{k+1/2}) - F(\mathbf{z}_{k+1/2})\big),$$

where by definition we set $\mathcal{F}_0 = \sigma(\xi_0, \xi_{-1}, \mathbf{w}_0) = \sigma(\xi_0)$. With this, we obtain the bound

$$\begin{aligned} \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} e_1(\mathbf{z}, k)\right] &= \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \langle \mathbf{u}_{k+1}, \mathbf{z}\rangle\right] - \mathbb{E}\left[\sum_{k=0}^{K-1} \langle \mathbf{u}_{k+1}, \mathbf{z}_{k+1/2}\rangle\right] \\ &= \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{k=0}^{K-1} \langle \mathbf{u}_{k+1}, \mathbf{z}\rangle\right] \\ &\le \max_{\mathbf{z}\in\mathcal{C}} \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 + \frac{1}{2}\sum_{k=0}^{K-1} \mathbb{E}\,\|\mathbf{u}_{k+1}\|^2 \\ &\le \max_{z\in\mathcal{C}} \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 + 2\tau^2 L^2 \sum_{k=0}^{K-1} \mathbb{E}\,\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2, \end{aligned} \tag{23}$$

where the second equality follows by the tower property, $\mathbb{E}_k[\mathbf{u}_{k+1}] = 0$, and $\mathcal{F}_k$-measurability of $\mathbf{z}_{k+1/2}$. The last inequality is due to

$$\begin{aligned}\mathbb{E}\,\|\mathbf{u}_{k+1}\|^2 &= \mathbb{E}\left[\mathbb{E}_k\,\|\mathbf{u}_{k+1}\|^2\right]\\ &= 4\tau^2\,\mathbb{E}\left[\mathbb{E}_k\,\|[F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k)] - F(\mathbf{z}_{k+1/2}) - F(\mathbf{w}_k)\|^2\right]\\ &\le 4\tau^2\,\mathbb{E}\left[\mathbb{E}_k\,\|F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k)\|^2\right]\\ &\le 4\tau^2 L^2\,\mathbb{E}\,\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2,\end{aligned}$$

where we use the tower property, $\mathbb{E}\,\|X - \mathbb{E}\,X\|^2 \le \mathbb{E}\,\|X\|^2$, and Assumption 1(iv).

Secondly, we set in Lemma 3

$$\mathcal{F}_k = \sigma(\xi_0, \ldots, \xi_k, \mathbf{w}_k), \qquad \tilde{\mathbf{x}}_0 = \mathbf{z}_0, \qquad \mathbf{u}_{k+1} = p\mathbf{z}_{k+1} + (1-p)\mathbf{w}_k - \mathbf{w}_{k+1},$$

and use $\mathbb{E}\left[\mathbb{E}_{k+1/2}[\|\mathbf{w}_{k+1}\|^2 - p\|\mathbf{z}_{k+1}\|^2 - (1-p)\|\mathbf{w}_k\|^2]\right] = 0$, to obtain the bound

$$\begin{aligned}\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1} e_2(\mathbf{z}, k)\right] &= 2\,\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\langle\mathbf{u}_{k+1}, z\rangle\right]\\ &\le \max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0 - \mathbf{z}\|^2 + \sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{u}_{k+1}\|^2\\ &= \max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0 - \mathbf{z}\|^2 + p(1-p)\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2, \qquad (24)\end{aligned}$$

where the inequality follows from Lemma 3 and the second equality from the derivation

$$\begin{aligned}\mathbb{E}\,\|\mathbf{u}_{k+1}\|^2 &= \mathbb{E}\left[\mathbb{E}_{k+1/2}\,\|\mathbf{u}_{k+1}\|^2\right] = \mathbb{E}\left[\mathbb{E}_{k+1/2}\,\|\,\mathbb{E}_{k+1/2}\,[\mathbf{w}_{k+1}] - \mathbf{w}_{k+1}\|^2\right]\\ &= \mathbb{E}\left[\mathbb{E}_{k+1/2}\,\|\mathbf{w}_{k+1}\|^2 - \|\,\mathbb{E}_{k+1/2}\,[\mathbf{w}_{k+1}]\,\|^2\right]\\ &= \mathbb{E}\left[p\|\mathbf{z}_{k+1}\|^2 + (1-p)\|\mathbf{w}_k\|^2 - \|p\mathbf{z}_{k+1} + (1-p)\mathbf{w}_k\|^2\right]\\ &= p(1-p)\,\mathbb{E}\,\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2,\end{aligned}$$

which uses $\mathbb{E}\,\|X - \mathbb{E}\,X\|^2 = \mathbb{E}\,\|X\|^2 - \|\,\mathbb{E}\,X\|^2$.

Combining (23), (24), and (22), we finally arrive at

$$\begin{aligned}2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \le \max_{\mathbf{z}\in\mathcal{C}}\Phi_0(\mathbf{z}) + \max_{\mathbf{z}\in\mathcal{C}}\frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 + 2\tau^2L^2\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2\\ + \max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0 - \mathbf{z}\|^2 + p(1-p)\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2 \qquad (25)\end{aligned}$$

We have to estimate terms under the sum:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{K-1}\left(2\tau^2L^2\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+p(1-p)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2\right)\right]\\
&\le p\,\mathbb{E}\left[\sum_{k=0}^{K-1}\left(2\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2\right)\right]\\
&\le p\,\mathbb{E}\left[\sum_{k=0}^{K-1}\left((2+\sqrt{2})\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+(2+\sqrt{2})\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right)\right]\\
&\le \frac{2+\sqrt{2}}{1-\gamma}\Phi_0(\mathbf{z}_*)\le\frac{3.5}{1-\gamma}\max_{\mathbf{z}\in\mathcal{C}}\Phi_0(\mathbf{z}),
\end{aligned}\tag{26}$$

where the first inequality in (26) uses Lemma 1 and $1-\alpha=p$.

Now we use $\mathbf{w}_0=\mathbf{z}_0$ and, hence, $\Phi_0(\mathbf{z})=(2-p)\|\mathbf{z}_0-\mathbf{z}\|^2\le 2\|\mathbf{z}_0-\mathbf{z}\|^2$ in (25). This yields

$$2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right]\le\left(2+\frac{3}{2}+\frac{7}{1-\gamma}\right)\max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2=7\left(\frac{1}{2}+\frac{1}{1-\gamma}\right)\max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2.$$

Finally, using $\tau=\frac{\sqrt{p}\gamma}{L}$, we obtain

$$\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right]\le\frac{7L}{2\sqrt{p}\gamma K}\left(\frac{1}{2}+\frac{1}{1-\gamma}\right)\max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2=\mathcal{O}\left(\frac{L}{\sqrt{p}K}\right).$$

With a stepsize $\tau=\frac{\sqrt{p}}{2L}$, the right-hand side reduces to $\frac{17.5L}{\sqrt{p}K}\max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2$. ■

**Proof of Corollary 6** In average each iteration costs $pN+2$ calls to $F_\xi$. To reach $\varepsilon$-accuracy we need $\left\lceil\mathcal{O}\left(\frac{L}{\sqrt{p}\varepsilon}\right)\right\rceil$ iterations. Hence, the total average complexity is $\mathcal{O}\left(\frac{(pN+2)L}{\sqrt{p}\varepsilon}\right)$. Finally, the optimal choice $p=\frac{2}{N}$ gives $\mathcal{O}\left(\frac{\sqrt{N}L}{\varepsilon}\right)$ complexity. ■

### A.2. Proof of Theorem 2

**Proof of Theorem 2** By the proof of Lemma 1, without removing the term $-\alpha\|\mathbf{z}_{k+1/2}-\mathbf{z}_k\|^2$ in (7), we have

$$\begin{aligned}
\mathbb{E}_k\left[\Phi_{k+1}(\mathbf{z}_*)\right]\le\Phi_k(\mathbf{z}_*)-(1-\gamma)\left((1-\alpha)\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right]\right)\\
-\alpha\|\mathbf{z}_{k+1/2}-\mathbf{z}_k\|^2.
\end{aligned}\tag{27}$$

By Robbins-Siegmund theorem (Robbins and Siegmund, 1971, Theorem 1), we have that $\Phi_k(\mathbf{z}_\star)$ converges a.s. and $\|\mathbf{z}_{k+1/2}-\mathbf{z}_k\|$, $\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|$ converges to 0 a.s.

Let $\mathsf{Z}_k=\begin{bmatrix}\mathbf{z}_k\\ \mathbf{w}_k\end{bmatrix}$ and $\mathsf{Z}_*=\begin{bmatrix}\mathbf{z}_*\\ \mathbf{z}_*\end{bmatrix}$, then $\Phi_k(\mathbf{z}_*)=\alpha\|\mathbf{z}_k-\mathbf{z}_*\|^2+\frac{1-\alpha}{p}\|\mathbf{w}_k-\mathbf{z}_*\|^2=\|\mathsf{Z}_k-\mathsf{Z}_*\|_Q^2$ with $Q=\mathrm{diag}(\alpha,\dots,\alpha,\frac{1-\alpha}{p},\dots,\frac{1-\alpha}{p})$. Then applying (Combettes and Pesquet, 2015, Proposition 2.3) to the inequality $\mathbb{E}_k\|\mathsf{Z}_{k+1}-\mathsf{Z}_*\|_Q^2\le\|\mathsf{Z}_k-\mathsf{Z}_*\|_Q^2$, we can construct $\Xi$, with

$\mathbb{P}(\Xi) = 1$, such that for all $\theta \in \Xi$ and $\forall \mathbf{z}_* \in \mathsf{Sol}$ $\|\mathsf{Z}_k(\theta) - \mathsf{Z}_*\|_Q$ converges and therefore, there exists $\Xi$ with $\mathbb{P}(\Xi) = 1$, such that

$$\forall \theta \in \Xi \text{ and } \forall \mathbf{z}_* \in \mathsf{Sol}, \quad \alpha\|\mathbf{z}_k(\theta) - \mathbf{z}_*\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k(\theta) - \mathbf{z}_*\|^2 \text{ converges.} \tag{28}$$

Moreover, by taking total expectation on (27), we get $\sum_{k=1}^{\infty} \mathbb{E}\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 < \infty$. By Fubini-Tonelli theorem, we have $\mathbb{E}\left[\sum_{k=1}^{\infty} \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2\right] < \infty$ and since $\sum_{k=1}^{\infty} \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2$ is nonnegative, $\sum_{k=1}^{\infty} \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 < \infty$ a.s. and thus $\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|$ converges to 0 a.s.

Let $\Xi'$ be the probability 1 set such that for all $\theta \in \Xi'$, $\mathbf{z}_{k+1}(\theta) - \mathbf{z}_{k+1/2}(\theta) \to 0$, $\mathbf{z}_{k+1/2}(\theta) - \mathbf{z}_k(\theta) \to 0$, and $\mathbf{z}_{k+1/2}(\theta) - \mathbf{w}_k(\theta) \to 0$. Pick $\theta \in \Xi \cap \Xi'$ and let $\tilde{\mathbf{z}}(\theta)$ be a cluster point of the bounded sequence $(\mathbf{z}_k(\theta))$. From $\mathbf{z}_{k+1/2}(\theta) - \mathbf{z}_k(\theta) \to 0$ and $\mathbf{z}_{k+1/2}(\theta) - \mathbf{w}_k(\theta) \to 0$ it follows that $\tilde{\mathbf{z}}(\theta)$ is also a cluster point of $(\mathbf{w}_k(\theta))$.

By prox-inequality (3) applied to the definition of $\mathbf{z}_{k+1}$,

$$\begin{aligned}\langle \mathbf{z}_{k+1}(\theta) - \bar{\mathbf{z}}_k(\theta) + \tau F(\mathbf{w}_k(\theta)) - \tau F_{\xi_k}(\mathbf{z}_{k+1/2}(\theta)) + \tau F_{\xi_k}(\mathbf{w}_k(\theta)), \mathbf{z} - \mathbf{z}_{k+1}(\theta)\rangle \\ + \tau g(\mathbf{z}) - \tau g(\mathbf{z}_{k+1}(\theta)) \geq 0, \quad \forall \mathbf{z} \in \mathcal{Z}.\end{aligned} \tag{29}$$

By extracting the subsequence of $\mathbf{z}_k(\theta)$ if needed, taking the limit along that subsequence and using the lower semicontinuity of $g$, we deduce that $\tilde{\mathbf{z}}(\theta) \in \mathsf{Sol}$. In doing so, we also used that $(\mathbf{z}_{k+1}(\theta))$ is bounded and $F_\xi$ is continuous for all $\xi$ to deduce $\tau\langle F_{\xi_k}(\mathbf{w}_k(\theta)) - F_{\xi_k}(\mathbf{z}_{k+1/2}(\theta)), \mathbf{z} - \mathbf{z}_{k+1}(\theta)\rangle \to 0$. Moreover, since $\mathbf{z}_{k+1}(\theta) - \mathbf{z}_k(\theta) \to 0$ and $\mathbf{z}_{k+1}(\theta) - \mathbf{w}_k(\theta) \to 0$, it follows that $\mathbf{z}_{k+1}(\theta) - \bar{\mathbf{z}}_k(\theta) \to 0$.

Hence, all cluster points of $(\mathbf{z}_k(\theta))$ and $(\mathbf{w}_k(\theta))$ belong to $\mathsf{Sol}$. We have shown that at least on one subsequence $\alpha\|\mathbf{z}_k(\theta) - \tilde{\mathbf{z}}(\theta)\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k(\theta) - \tilde{\mathbf{z}}(\theta)\|^2$ converges to 0. Then, by (28) we deduce $\alpha\|\mathbf{z}_k(\theta) - \tilde{\mathbf{z}}(\theta)\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k(\theta) - \tilde{\mathbf{z}}(\theta)\|^2 \to 0$ and consequently $\|\mathbf{z}_k(\theta) - \tilde{\mathbf{z}}(\theta)\|^2 \to 0$. This shows $(\mathbf{z}_k)$ converges almost surely to a point in $\mathsf{Sol}$. ■

## Appendix B. Analysis for Section 3

**Remark 12** *For running Alg. 2 in practice, we suggest* $K = \frac{N}{2}$, $\alpha = 1 - \frac{1}{K}$, *and* $\tau = \frac{0.99\sqrt{p}}{L}$.

We recall the three point identity which can be seen as the analogue of the standard Euclidean identity $2\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a} + \mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2$:

$$\langle \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}), \mathbf{z} - \mathbf{x}\rangle = D(\mathbf{z}, \mathbf{y}) - D(\mathbf{z}, \mathbf{x}) - D(\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{Z}. \tag{30}$$

Similar to Euclidean case, we define for the iterates $(\mathbf{z}_k^s)$ of Alg. 2 and any $\mathbf{z} \in \operatorname{dom} g$,

$$\Phi^s(\mathbf{z}) := \alpha D(\mathbf{z}, \mathbf{z}_0^s) + (1-\alpha)\sum_{j=1}^{K} D(\mathbf{z}, \mathbf{z}_j^{s-1}),$$

where $\Phi^0(\mathbf{z}) = (\alpha + K(1-\alpha))D(\mathbf{z}, \mathbf{z}_0)$, due to the definition of $\mathbf{z}^{-1}$ in Alg. 2. Since we have two indices $s, k$ in Alg. 2, we define $\mathcal{F}_k^s = \sigma(\mathbf{z}_{1/2}^0, \ldots, \mathbf{z}_{k-1/2}^0, \ldots, \mathbf{z}_{1/2}^s, \ldots, \mathbf{z}_{k+1/2}^s)$ and $\mathbb{E}_{s,k}[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k^s]$.

**Lemma 13** *Let $g$ be proper convex lsc and*

$$\mathbf{z}^+ = \operatorname*{argmin}_{\mathbf{z}} \{g(\mathbf{z}) + \langle \mathbf{u}, \mathbf{z}\rangle + \alpha D(\mathbf{z}, \mathbf{z}_1) + (1-\alpha) D(\mathbf{z}, \mathbf{z}_2)\}.$$

*Then, for any $\mathbf{z} \in \mathcal{Z}$,*

$$\begin{aligned} g(\mathbf{z}) - g(\mathbf{z}^+) + \langle \mathbf{u}, \mathbf{z} - \mathbf{z}^+\rangle \geq D(\mathbf{z}, \mathbf{z}^+) + \alpha \left(D(\mathbf{z}^+, \mathbf{z}_1) - D(\mathbf{z}, \mathbf{z}_1)\right) \\ + (1-\alpha)\left(D(\mathbf{z}^+, \mathbf{z}_2) - D(\mathbf{z}, \mathbf{z}_2)\right). \end{aligned}$$

**Proof** By optimality of $\mathbf{z}^+$,

$$0 \in \partial g(\mathbf{z}^+) + \mathbf{u} + \alpha\left(\nabla h(\mathbf{z}^+) - \nabla h(\mathbf{z}_1)\right) + (1-\alpha)\left(\nabla h(\mathbf{z}^+) - \nabla h(\mathbf{z}_2)\right).$$

This implies by convexity of $g$

$$g(\mathbf{z}) - g(\mathbf{z}^+) \geq \langle \mathbf{u} + \alpha\left(\nabla h(\mathbf{z}^+) - \nabla h(\mathbf{z}_1)\right) + (1-\alpha)\left(\nabla h(\mathbf{z}^+) - \nabla h(\mathbf{z}_2)\right), \mathbf{z}^+ - \mathbf{z}\rangle.$$

By applying three point identity twice, we deduce

$$\begin{aligned} g(\mathbf{z}) - g(\mathbf{z}^+) + \langle \mathbf{u}, \mathbf{z} - \mathbf{z}^+\rangle \geq \alpha\left(D(\mathbf{z}, \mathbf{z}^+) + D(\mathbf{z}^+, \mathbf{z}_1) - D(\mathbf{z}, \mathbf{z}_1)\right) \\ + (1-\alpha)\left(D(\mathbf{z}, \mathbf{z}^+) + D(\mathbf{z}^+, \mathbf{z}_2) - D(\mathbf{z}, \mathbf{z}_2)\right) \end{aligned}$$

and by a simple rearrangement we obtain the result. ■

We now introduce some definitions to be used in the proofs of this section.

$$\Theta^s_{k+1/2}(\mathbf{z}) = \langle F(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1/2} - \mathbf{z}\rangle + g(\mathbf{z}^s_{k+1/2}) - g(\mathbf{z}), \tag{31}$$

$$e(\mathbf{z}, s, k) = \tau\langle F(\mathbf{z}^s_{k+1/2}) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2}) - F(\mathbf{w}^s) + F_{\xi^s_k}(\mathbf{w}^s), \mathbf{z}^s_{k+1/2} - \mathbf{z}\rangle. \tag{32}$$

$$\begin{aligned} \delta(s,k) = \tau\langle F_{\xi^s_k}(\mathbf{w}^s) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\rangle - \frac{1}{2}\|\mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\|^2 \\ - \frac{1-\alpha}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2 \end{aligned} \tag{33}$$

The first expression will be needed for deriving the rate, the second term $e(\mathbf{z}, s, k)$ for controlling the error caused by $\max_{\mathbf{z}\in\mathcal{C}} \mathbb{E}[\cdot] \neq \mathbb{E}\max_{\mathbf{z}\in\mathcal{C}}[\cdot]$, and the third term $\delta(s,k)$ will be nonpositive after taking expectation.

**Lemma 14** *Let Assumption 1 hold, $\alpha \in [0,1)$, and $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ for $\gamma \in (0,1)$. Then we have the following:*

*(i) For any $\mathbf{z} \in \mathcal{Z}$ and $s, K \in \mathbb{N}$, it holds that*

$$\begin{aligned} \sum_{k=0}^{K-1} \tau\Theta^s_{k+1/2}(\mathbf{z}) + \alpha D(\mathbf{z}, \mathbf{z}_0^{s+1}) + (1-\alpha)\sum_{j=1}^{K} D(\mathbf{z}, \mathbf{z}_j^s) \\ \leq \alpha D(\mathbf{z}, \mathbf{z}_0^s) + (1-\alpha)\sum_{j=1}^{K} D(\mathbf{z}, \mathbf{z}_j^{s-1}) + \sum_{k=0}^{K-1}[e(\mathbf{z}, s, k) + \delta(s,k)]. \end{aligned}$$

*(ii) For any solution $\mathbf{z}_*$, it holds that*

$$\mathbb{E}_{s,0}\left[\Phi^{s+1}(\mathbf{z}_*)\right] \le \Phi^s(\mathbf{z}_*) - \frac{(1-\alpha)(1-\gamma^2)}{2}\sum_{k=0}^{K-1}\mathbb{E}_{s,0}\left[\|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2\right].$$

*(iii) It holds that* $\sum_{s=0}^{\infty}\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2 \le \frac{2}{(1-\alpha)(1-\gamma^2)}\Phi^0(\mathbf{z}_*)$.

**Remark 15** *We use Lemma 14(i) and Lemma 14(iii) for proving the convergence rate. Moreover, Lemma 14(ii) can be used to derive subsequential convergence, which we do not include for brevity.*

**Proof of Lemma 14** Applying Lemma 13 to $\mathbf{z}_{k+1/2}^s$ update, with $\mathbf{z} = \mathbf{z}_{k+1}^s$, we have

$$\begin{aligned}\tau\left(g(\mathbf{z}_{k+1}^s) - g(\mathbf{z}_{k+1/2}^s) + \langle F(\mathbf{w}^s), \mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\rangle\right) &\ge D(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s)\\ &+ \alpha\Big(D(\mathbf{z}_{k+1/2}^s, \mathbf{z}_k^s) - D(\mathbf{z}_{k+1}^s, \mathbf{z}_k^s)\Big) + (1-\alpha)\Big(D(\mathbf{z}_{k+1/2}^s, \bar{\mathbf{w}}^s) - D(\mathbf{z}_{k+1}^s, \bar{\mathbf{w}}^s)\Big). \quad (34)\end{aligned}$$

Applying Lemma 13 to $\mathbf{z}_{k+1}^s$ update with a general $\mathbf{z} \in \mathcal{Z}$, we have

$$\begin{aligned}\tau\left(g(\mathbf{z}) - g(\mathbf{z}_{k+1}^s) + \langle \widehat{F}(\mathbf{z}_{k+1/2}^s), \mathbf{z} - \mathbf{z}_{k+1}^s\rangle\right) &\ge D(\mathbf{z}, \mathbf{z}_{k+1}^s)\\ &+ \alpha\Big(D(\mathbf{z}_{k+1}^s, \mathbf{z}_k^s) - D(\mathbf{z}, \mathbf{z}_k^s)\Big) + (1-\alpha)\Big(D(\mathbf{z}_{k+1}^s, \bar{\mathbf{w}}^s) - D(\mathbf{z}, \bar{\mathbf{w}}^s)\Big). \quad (35)\end{aligned}$$

Note that for any $\mathbf{u}, \mathbf{v}$, the expression $D(\mathbf{u}, \bar{\mathbf{w}}^s) - D(\mathbf{v}, \bar{\mathbf{w}}^s)$ is linear in terms of $\nabla h(\bar{\mathbf{w}}^s)$, that is

$$D(\mathbf{u}, \bar{\mathbf{w}}^s) - D(\mathbf{v}, \bar{\mathbf{w}}^s) = \frac{1}{K}\sum_{j=1}^{K}\Big(D(\mathbf{u}, \mathbf{z}_j^{s-1}) - D(\mathbf{v}, \mathbf{z}_j^{s-1})\Big). \quad (36)$$

Summing up (34) and (35) and using (36) with definition of $\widehat{F}(\mathbf{z}_{k+1/2}^s)$, we obtain

$$\begin{aligned}\tau\Big(g(\mathbf{z}) - g(\mathbf{z}_{k+1/2}^s) + \langle \widehat{F}(\mathbf{z}_{k+1/2}^s), \mathbf{z} - \mathbf{z}_{k+1/2}^s\rangle\Big) &\ge D(\mathbf{z}, \mathbf{z}_{k+1}^s) - \alpha D(\mathbf{z}, \mathbf{z}_k^s)\\ &+ \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}_{k+1/2}^s, \mathbf{z}_j^{s-1}) - \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}, \mathbf{z}_j^{s-1}) + D(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s)\\ &+ \tau\langle F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s), \mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\rangle. \quad (37)\end{aligned}$$

By $D(\mathbf{u}, \mathbf{v}) \ge \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2$ and Jensen's inequality, we have

$$\begin{aligned}\frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}_{k+1/2}^s, \mathbf{z}_j^{s-1}) &\ge \frac{1-\alpha}{K}\sum_{j=1}^{K}\frac{1}{2}\|\mathbf{z}_{k+1/2}^s - \mathbf{z}_j^{s-1}\|^2 \ge \frac{1-\alpha}{2}\|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2, \quad (38)\\ D(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) &\ge \frac{1}{2}\|\mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\|^2. \quad (39)\end{aligned}$$

By using (31), (38), and (39) in (37), we deduce

$$\tau\Theta^s_{k+1/2}(\mathbf{z}) + D(\mathbf{z}, \mathbf{z}^s_{k+1}) \le \alpha D(\mathbf{z}, \mathbf{z}^s_k) + \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}, \mathbf{z}^{s-1}_j)$$
$$+ \tau\langle F_{\xi^s_k}(\mathbf{w}^s) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\rangle - \frac{1}{2}\|\mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\|^2 - \frac{1-\alpha}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2$$
$$+ \underbrace{\tau\langle F(\mathbf{z}^s_{k+1/2}) - \widehat{F}(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1/2} - \mathbf{z}\rangle}_{e(\mathbf{z},s,k)},$$

where we defined the last term as $e(\mathbf{z}, s, k)$ (see (32)). We sum this inequality over $k$ to obtain the result in $(i)$.

Next, similar to (9), we estimate by Assumption 2 and Young's inequality

$$\tau\mathbb{E}_{s,k}\langle F_{\xi^s_k}(\mathbf{w}^s) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\rangle$$
$$\le \mathbb{E}_{s,k}\left[\frac{\tau^2}{2}\|F_{\xi^s_k}(\mathbf{w}^s) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2})\|^2_* + \frac{1}{2}\|\mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\|^2\right]$$
$$\le \frac{(1-\alpha)\gamma^2}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2 + \frac{1}{2}\mathbb{E}_{s,k}\|\mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\|^2, \quad (40)$$

since $\tau^2 L^2 = (1-\alpha)\gamma^2$. We take expectation of (37), plug in $\mathbf{z} = \mathbf{z}_*$; use (8), (40), (38), and (39) to get

$$\mathbb{E}_{s,k}\left[D(\mathbf{z}_*, \mathbf{z}^s_{k+1})\right] \le \alpha D(\mathbf{z}_*, \mathbf{z}^s_k) + \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}_*, \mathbf{z}^{s-1}_j)$$
$$+ \frac{(1-\alpha)(\gamma^2-1)}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2. \quad (41)$$

By using $\mathbb{E}_{s,0}[\cdot] = \mathbb{E}_{s,0}\left[\mathbb{E}_{s,k}[\cdot]\right]$, we have

$$\mathbb{E}_{s,0}D(\mathbf{z}_*, \mathbf{z}^s_{k+1}) \le \mathbb{E}_{s,0}\Big[\alpha D(\mathbf{z}_*, \mathbf{z}^s_k) + \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}_*, \mathbf{z}^{s-1}_j)$$
$$- \frac{(1-\alpha)(1-\gamma^2)}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2\Big]. \quad (42)$$

Summing this inequality over $k = 0, \dots, K-1$ and using the definition of $\Phi^s(\mathbf{z}_*)$ together with $\mathbf{z}^{s+1}_0 = \mathbf{z}^s_K$, we derive *(ii)*.

Finally, we take total expectation of *(ii)* and sum the inequality over $s$ to obtain *(iii)*. ■

In order to prove the convergence rate, we need the Bregman version of Lemma 3.

**Lemma 16** *Let $\mathcal{F} = (\mathcal{F}^s_k)_{s\ge 0, k\in[0,K-1]}$ be a filtration and $(\mathbf{u}^s_k)$ a stochastic process adapted to $\mathcal{F}$ with $\mathbb{E}[\mathbf{u}^s_{k+1}|\mathcal{F}^s_k] = 0$. Given $\mathbf{x}_0 \in \mathcal{Z}$, for any $S \in \mathbb{N}$ and any compact set $\mathcal{C} \subset \operatorname{dom} g$*

$$\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{C}}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\langle \mathbf{u}^s_{k+1}, \mathbf{x}\rangle\right] \le \max_{\mathbf{x}\in\mathcal{C}} D(\mathbf{x}, \mathbf{x}_0) + \frac{1}{2}\sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{u}^s_{k+1}\|^2_*.$$

**Proof** Define for each $s \geq 0$ and for $k \in \{0, \dots, K-1\}$,

$$\mathbf{x}_{k+1}^s = \operatorname*{argmin}_{\mathbf{x} \in \operatorname{dom} g} \{\langle -\mathbf{u}_{k+1}^s, \mathbf{x} \rangle + D(\mathbf{x}, \mathbf{x}_k^s)\}, \text{ and let } \mathbf{x}_0^{s+1} = \mathbf{x}_m^s.$$

First, we observe $\mathbf{x}_k^s$ is $\mathcal{F}_k^s$-measurable. By the definition of $\mathbf{x}_{k+1}^s$, we have for all $\mathbf{x} \in \operatorname{dom} g$,

$$\langle \nabla h(\mathbf{x}_{k+1}^s) - \nabla h(\mathbf{x}_k^s) - \mathbf{u}_{k+1}^s, \mathbf{x} - \mathbf{x}_{k+1}^s \rangle \geq 0.$$

We apply three point identity to obtain

$$D(\mathbf{x}, \mathbf{x}_k^s) - D(\mathbf{x}, \mathbf{x}_{k+1}^s) - D(\mathbf{x}_{k+1}^s, \mathbf{x}_k^s) - \langle \mathbf{u}_{k+1}^s, \mathbf{x} - \mathbf{x}_{k+1}^s \rangle \geq 0.$$

We bound the inner product by using Hölder's, Young's inequalities, and strong convexity of $h$,

$$\begin{aligned}
\langle \mathbf{u}_{k+1}^s, \mathbf{x} - \mathbf{x}_{k+1}^s \rangle &= \langle \mathbf{u}_{k+1}^s, \mathbf{x} - \mathbf{x}_k^s \rangle + \langle \mathbf{u}_{k+1}^s, \mathbf{x}_k^s - \mathbf{x}_{k+1}^s \rangle \\
&\geq \langle \mathbf{u}_{k+1}^s, \mathbf{x} - \mathbf{x}_k^s \rangle - \frac{1}{2}\|\mathbf{u}_{k+1}^s\|_*^2 - \frac{1}{2}\|\mathbf{x}_{k+1}^s - \mathbf{x}_k^s\|^2 \\
&\geq \langle \mathbf{u}_{k+1}^s, \mathbf{x} - \mathbf{x}_k^s \rangle - \frac{1}{2}\|\mathbf{u}_{k+1}^s\|_*^2 - D(\mathbf{x}_{k+1}^s, \mathbf{x}_k^s),
\end{aligned}$$

which, combined with the previous inequality gives

$$\langle \mathbf{u}_{k+1}^s, \mathbf{x} \rangle \leq D(\mathbf{x}, \mathbf{x}_k^s) - D(\mathbf{x}, \mathbf{x}_{k+1}^s) + \langle \mathbf{u}_{k+1}^s, \mathbf{x}_k^s \rangle + \frac{1}{2}\|\mathbf{u}_{k+1}^s\|_*^2. \tag{43}$$

We sum (43) over $k = 0, \dots, K-1$ and $s = 0, \dots, S-1$, take maximum, use $\mathbf{x}_0^{s+1} = \mathbf{x}_K^s$ and the same derivations as the end of the proof of Lemma 3 to show $\sum_{s=0}^{S-1} \sum_{k=0}^{K-1} \mathbb{E}\left[\langle \mathbf{u}_{k+1}^s, \mathbf{x}_k^s \rangle\right] = 0$. ■

**Proof of Theorem 8** We start with the result of Lemma 14 and proceed similar to Theorem 4. Since $\mathbf{z}_0^{s+1} = \mathbf{z}_K^s$, we use definition of $\Phi^s(\mathbf{z})$, and sum the inequality in Lemma 14(i) over $s$ to obtain

$$\sum_{s=0}^{S-1} \sum_{k=0}^{K-1} \tau \Theta_{k+1/2}^s(\mathbf{z}) + \Phi^S(\mathbf{z}) \leq \Phi^0(\mathbf{z}) + \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} [e(\mathbf{z}, s, k) + \delta(s, k)]$$

We take maximum and expectation, use $\mathbb{E}\left[\max_{\mathbf{z} \in \mathcal{C}} \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} \tau \Theta_{k+1/2}^s(\mathbf{z})\right] \geq \tau K S \mathbb{E}\left[\operatorname{Gap}(\mathbf{z}^S)\right]$ to deduce

$$\tau K S \mathbb{E}\left[\operatorname{Gap}(\mathbf{z}^S)\right] \leq \max_{\mathbf{z} \in \mathcal{C}} \Phi^0(\mathbf{z}) + \mathbb{E}\left[\max_{\mathbf{z} \in \mathcal{C}} \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} e(\mathbf{z}, s, k)\right] + \mathbb{E}\left[\sum_{s=0}^{S-1} \sum_{k=0}^{K-1} \delta(s, k)\right].$$

The term $\mathbb{E} \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} \delta(s, k)$ is nonpositive by the tower property, Lipschitzness, Young's inequality, and $\tau < \frac{\sqrt{p}}{L}$ (the same arguments used in (40) can be applied here with $\delta(s, k)$ defined as (33)). Therefore,

$$\tau K S \mathbb{E}\left[\operatorname{Gap}(\mathbf{z}^S)\right] \leq \max_{\mathbf{z} \in \mathcal{C}} \Phi^0(\mathbf{z}) + \mathbb{E}\left[\max_{\mathbf{z} \in \mathcal{C}} \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} e(\mathbf{z}, s, k)\right].$$

We bound the second term on RHS, similar to the proof of Theorem 4. For $s \in \{0, \dots, S-1\}$ and $k \in \{0, \dots, K-1\}$, set $\mathcal{F}_k^s = \sigma(\mathbf{z}_{1/2}^0, \dots, \mathbf{z}_{K-1/2}^0, \dots, \mathbf{z}_{1/2}^s, \dots, \mathbf{z}_{k+1/2}^s)$, $\mathbf{u}_{k+1}^s = \tau[\widehat{F}(\mathbf{z}_{k+1/2}^s) - F(\mathbf{z}_{k+1/2}^s)] = \tau[F(\mathbf{w}^s) - F_{\xi_k^s}(\mathbf{w}^s) - F(\mathbf{z}_{k+1/2}^s) + F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s)]$, which help us write

$$\begin{aligned}
\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} e(\mathbf{z},k)\right] &= \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \tau\langle \widehat{F}(\mathbf{z}_{k+1/2}^s) - F(\mathbf{z}_{k+1/2}^s), \mathbf{z} - \mathbf{z}_{k+1/2}^s\rangle\right] \\
&= \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \langle \mathbf{u}_{k+1}^s, \mathbf{z}\rangle\right] - \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \mathbb{E}\langle \mathbf{u}_{k+1}^s, \mathbf{z}_{k+1/2}^s\rangle \\
&= \mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \langle \mathbf{u}_{k+1}^s, \mathbf{z}\rangle\right],
\end{aligned}$$

where the last equality is due to the tower property, $\mathcal{F}_k^s$-measurability of $\mathbf{z}_{k+1/2}^s$ and $\mathbb{E}_{s,k}[\mathbf{u}_{k+1}^s] = 0$.

We apply Lemma 16 with the specified $\mathcal{F}_k^s$, $\mathbf{u}_{k+1}^s$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} e(\mathbf{z},k)\right] \\
&\qquad \le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \tau^2 \mathbb{E}\|F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s) + F(\mathbf{w}^s) - F(\mathbf{z}_{k+1/2}^s)\|_*^2 \\
&\qquad \le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} 4\tau^2 \mathbb{E}\|F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s)\|_*^2 \qquad (44) \\
&\qquad \le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} 4\tau^2 L^2 \mathbb{E}\|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2 \qquad (45) \\
&\qquad \le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \frac{8\tau^2 L^2}{(1-\alpha)(1-\gamma^2)} \Phi^0(\mathbf{z}_*), \qquad (46)
\end{aligned}$$

where (44) is due to the tower property and $\mathbb{E}\|X - \mathbb{E}X\|_*^2 \le 2\mathbb{E}\|X\|_*^2 + 2\|\mathbb{E}X\|_*^2 \le 4\mathbb{E}\|X\|_*^2$, which follows from triangle inequality, Young's inequality, and Jensen's inequality. Moreover, (45) is by variable Lipschitzness of $F_\xi$, and the last step is by Lemma 14. Consequently, by $\Phi^0(\mathbf{z}_*) \le \max_{\mathbf{z}\in\mathcal{C}} \Phi^0(\mathbf{z}) = (\alpha + K(1-\alpha)) \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0)$ and $\tau^2 L^2 = (1-\alpha)\gamma^2$ we have

$$\begin{aligned}
\tau K S \mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^S)\right] &\le \max_{\mathbf{z}\in\mathcal{C}} \left[D(\mathbf{z},\mathbf{z}_0) + \left(1 + \frac{8\tau^2 L^2}{(1-\alpha)(1-\gamma^2)}\right)\Phi^0(\mathbf{z})\right] \\
&= \left(1 + \left(1 + \frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha)\right) \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0),
\end{aligned}$$

which gives the result. ■

**Proof of Corollary 9** As $\alpha = 1 - \frac{1}{K}$, it holds that $\alpha + K(1-\alpha) = 1 - \frac{1}{K} + 1 \leq 2$. With this, from Theorem 8 it follows

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^S)\right] &\leq \frac{1}{\tau K S}\left(1 + \left(1 + \frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha)\right) \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0) \\ &\leq \frac{L}{\sqrt{K}\gamma S}\left(3 + \frac{16\gamma^2}{1-\gamma^2}\right) \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0) = \mathcal{O}\left(\frac{L}{\sqrt{N}S}\right). \end{aligned} \tag{47}$$

One epoch requires one evaluation of $F$ and $2K$ of $F_\xi$, therefore in total we have $N + 2K = 2N$. To reach $\varepsilon$ accuracy, we need $\left\lceil \mathcal{O}\left(\frac{L}{\sqrt{N}\varepsilon}\right)\right\rceil$ epochs. Hence, the final complexity is $\mathcal{O}\left(N + \frac{L\sqrt{N}}{\varepsilon}\right)$. Now, by setting $\gamma = \frac{1}{3}$ in (47), we will get specific constants. In particular, we will have

$$\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^S)\right] \leq \frac{15L}{\sqrt{K}S} \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0) = \frac{15\sqrt{2}L}{\sqrt{N}S} \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0).$$

Consequently, since $30\sqrt{2} < 43$, the final complexity is $\left(2N + \frac{43\sqrt{N}L}{\varepsilon} \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0)\right)$. ■

**Remark 17** *Because we work with general norms, we had to use in (46) a crude inequality* $\mathbb{E}\|X - \mathbb{E}X\|_*^2 \leq 4\mathbb{E}\|X\|_*^2$. *Of course, in the Euclidean case with* $D(\mathbf{z}, \mathbf{z}') = \frac{1}{2}\|\mathbf{z} - \mathbf{z}'\|^2$ *this factor* 4 *is redundant. It is easy to see that setting* $\tau = \frac{\sqrt{1-\alpha}}{2L}$ *and the rest of the parameters as in Corollary 9 leads to* $\left(2N + \frac{13\sqrt{N}L}{\varepsilon} \max_{\mathbf{z}\in\mathcal{C}} \|\mathbf{z} - \mathbf{z}_0\|^2\right)$ *total complexity for the Euclidean setting.*

## Appendix C. Extensions

In this section, we show how to obtain the variance reduced versions of two other operator splitting methods: forward-backward-forward (FBF) (Tseng, 2000) and forward-reflected-backward (FoRB) (Malitsky and Tam, 2020) for monotone inclusions. We also show how to obtain linear convergence with Algorithm 1 when $g$ in (1) is strongly convex.

Formally, the monotone inclusion problem is to find

$$\mathbf{z}_* \in \mathcal{Z} \text{ such that } 0 \in (F + G)(\mathbf{z}_*), \tag{48}$$

where $\mathcal{Z}$ is a finite dimensional vector space with Euclidean inner product and the rest of the assumptions are summarized in Assumption 3.

**Assumption 3**

*(i) The solution set* Sol *of (48) is nonempty:* $(F+G)^{-1}(0) \neq \varnothing$.

*(ii) The operators* $G\colon \mathcal{Z} \rightrightarrows \mathcal{Z}$ *and* $F\colon \mathcal{Z} \to \mathcal{Z}$ *are maximally monotone.*

*(iii) The operator* $F$ *has an oracle* $F_\xi$ *that is unbiased* $F(\mathbf{z}) = \mathbb{E}_\xi\left[F_\xi(\mathbf{z})\right]$ *and* $L$*-Lipschitz in mean:*

$$\mathbb{E}_\xi\left[\|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|^2\right] \leq L^2\|\mathbf{u} - \mathbf{v}\|^2, \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{Z}.$$

**Algorithm 3** FBF with variance reduction

**Input:** Probability $p \in (0,1]$, probability distribution $Q$, step size $\tau$, $\alpha \in (0,1)$. Let $\mathbf{z}_0 = \mathbf{w}_0$
**for** $k = 0, 1 \dots$ **do**
  $\bar{\mathbf{z}}_k = \alpha \mathbf{z}_k + (1-\alpha)\mathbf{w}_k$
  $\mathbf{z}_{k+1/2} = J_{\tau G}(\bar{\mathbf{z}}_k - \tau F(\mathbf{w}_k))$
  Draw an index $\xi_k$ according to $Q$
  $\mathbf{z}_{k+1} = \mathbf{z}_{k+1/2} - \tau(F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k))$
  $\mathbf{w}_{k+1} = \begin{cases} \mathbf{z}_{k+1}, & \text{with probability } p \\ \mathbf{w}_k, & \text{with probability } 1-p \end{cases}$
**end for**

We remark that one can use variable Lipschitz assumption from Assumption 2 instead of standard Lipschitzness, but we chose the latter for simplicity. Let us also recall the conditional expectation definitions based on the iterates of the algorithms: $\mathbb{E}[\cdot|\sigma(\xi_0, \dots, \xi_{k-1}, \mathbf{w}_k)] = \mathbb{E}_k[\cdot]$ and $\mathbb{E}[\cdot|\sigma(\xi_0, \dots, \xi_k, \mathbf{w}_k] = \mathbb{E}_{k+1/2}[\cdot]$. Next, the resolvent of an operator $G$ is given by $J_G = (I+G)^{-1}$ where $I$ is the identity operator. It is easy to see that when $G = \partial g$ for proper convex lsc function $g$, inclusion (48) becomes the VI in (1) and $J_G = \operatorname{prox}_g$.

### C.1. Forward-Backward-Forward with variance reduction

Forward-backward-forward (FBF) algorithm was introduced by Tseng in (Tseng, 2000). On one hand, it is a modification of the forward-backward algorithm that does not require stronger assumptions than mere monotonicity. On the other, it is a modification of the extragradient method that works for general monotone inclusions and not just for variational inequalities. FBF reads as

$$\begin{cases} \mathbf{z}_{k+1/2} = J_{\tau G}(\mathbf{z}_k - \tau F(\mathbf{z}_k)) \\ \mathbf{z}_{k+1} = \mathbf{z}_{k+1/2} - \tau F(\mathbf{z}_{k+1/2}) + \tau F(\mathbf{z}_k). \end{cases}$$

It is easy to see that FBF is equivalent to extragradient when $G$ is absent. But when not, FBF applied to the VI requires one proximal operator every iteration, whereas extragradient requires two. This advantage can be important for the cases where proximal operator is computationally expensive (Böhm et al., 2020).

**Remark 18** *For running Alg. 3 in practice, we suggest* $p = \frac{2}{N}$, $\alpha = 1-p$, *and* $\tau = \frac{0.99\sqrt{p}}{L}$.

We keep the same notation as Section 2.3 and recall the definition of $\Phi_k$ for convenience

$$\Phi_k(\mathbf{z}) = \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k - \mathbf{z}\|^2.$$

We now continue with the main result for FBF.

**Theorem 19** *Let Assumption 3 hold,* $\alpha \in [0,1)$, $p \in (0,1]$, *and* $\tau = \frac{\sqrt{1-\alpha}}{L}\gamma$ *for* $\gamma \in (0,1)$. *Then for* $(\mathbf{z}_k)$ *generated by Alg. 3 and any* $\mathbf{z}_* \in \mathsf{Sol}$, *it holds that*

$$\mathbb{E}_k\left[\Phi_{k+1}(\mathbf{z}_*)\right] \le \Phi_k(\mathbf{z}_*).$$

*Moreover, if* $F_\xi$ *is continuous for all* $\xi$, *then* $(\mathbf{z}_k)$ *converges to some* $\mathbf{z}_* \in \mathsf{Sol}$ *a.s.*

**Proof** Let $\mathbf{z} = \mathbf{z}_* \in \mathsf{Sol}$ which gives $-F(\mathbf{z}) \in G(\mathbf{z})$. Next, by the definition of $\mathbf{z}_{k+1/2}$ and resolvent, $\bar{\mathbf{z}}_k - \tau F(\mathbf{w}_k) \in \mathbf{z}_{k+1/2} + \tau G(\mathbf{z}_{k+1/2})$. Combining these estimates with monotonicity of $G$ lead to

$$\langle \mathbf{z}_{k+1/2} - \bar{\mathbf{z}}_k + \tau F(\mathbf{w}_k), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle - \tau\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle \geq 0.$$

We plug in the definition of $\mathbf{z}_{k+1}$ into this inequality to obtain

$$\langle \mathbf{z}_{k+1} - \bar{\mathbf{z}}_k + \tau\left(F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k) + F(\mathbf{w}_k)\right), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle - \langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle \geq 0. \tag{49}$$

We estimate the term with $\bar{\mathbf{z}}_k$ as in (6)

$$\begin{aligned}
2\langle \mathbf{z}_{k+1} - \bar{\mathbf{z}}_k, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle &= 2\langle \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle + 2\langle \mathbf{z}_{k+1/2} - \bar{\mathbf{z}}_k, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle\\
&= \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 + \|\mathbf{z} - \mathbf{z}_{k+1/2}\|^2 - \|\mathbf{z} - \mathbf{z}_{k+1}\|^2 + 2\langle \mathbf{z}_{k+1/2} - \bar{\mathbf{z}}_k, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle\\
&= \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 - \|\mathbf{z} - \mathbf{z}_{k+1}\|^2 + \alpha\|\mathbf{z} - \mathbf{z}_k\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}\|^2\\
&\quad - \alpha\|\mathbf{z}_{k+1/2} - \mathbf{z}_k\|^2 - (1-\alpha)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2.
\end{aligned} \tag{50}$$

By taking conditional expectation and using that $\mathbf{z}_{k+1/2}$ is $\mathcal{F}_k$-measurable, we deduce

$$2\tau\mathbb{E}_k\left[\langle F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k) + F(\mathbf{w}_k), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle\right] = 2\tau\mathbb{E}_k\left[\langle F(\mathbf{z}_{k+1/2}), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle\right]. \tag{51}$$

We use (50) and (51) in (49) to obtain

$$\begin{aligned}
2\tau\langle F(\mathbf{z}) - F(\mathbf{z}_{k+1/2}), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 &\leq \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}\|^2\\
&\quad + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 - \alpha\|\mathbf{z}_{k+1/2} - \mathbf{z}_k\|^2 - (1-\alpha)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2.
\end{aligned}$$

Note that, the first term in the LHS is nonnegative by monotonicity of $F$. Then we add (11) to this inequality and use $\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 \leq \tau^2 L^2\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2$ to obtain

$$\begin{aligned}
\alpha\mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_{k+1} - \mathbf{z}\|^2 &\leq \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k - \mathbf{z}\|^2 - \alpha\|\mathbf{z}_{k+1/2} - \mathbf{z}_k\|^2\\
&\quad - \left((1-\alpha) - \tau^2L^2\right)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2.
\end{aligned}$$

This derives the first result, which is the analogue of Lemma 1. To show almost sure convergence, we basically follow the proof of Theorem 2. First, using Robbins-Siegmund theorem and (Combettes and Pesquet, 2015, Proposition 2.3) as in Theorem 2, we obtain that there exists a probability 1 set $\Xi$ of random trajectories such that $\forall\theta \in \Xi$ and $\forall\mathbf{z} \in \mathsf{Sol}$, we have that $\alpha\|\mathbf{z}_k(\theta) - \mathbf{z}\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k(\theta) - \mathbf{z}\|^2$ converges, $\mathbf{z}_{k+1/2}(\theta) - \mathbf{z}_k(\theta) \to 0$, and $\mathbf{z}_{k+1/2}(\theta) - \mathbf{w}_k(\theta) \to 0$. The latter implies $\mathbf{z}_{k+1}(\theta) - \mathbf{z}_{k+1/2}(\theta) \to 0$. Let $\tilde{\mathbf{z}}(\theta)$ be a cluster point of the bounded sequence $(\mathbf{z}_k(\theta))$. Instead of (29), we use the definitions of $\mathbf{z}_{k+1/2}$, resolvent, and $\mathbf{z}_{k+1}$ to obtain

$$\begin{aligned}
\mathbf{z}_{k+1}(\theta) - \bar{\mathbf{z}}_k(\theta) + \tau\left(F_{\xi_k}(\mathbf{w}_k(\theta)) - F_{\xi_k}(\mathbf{z}_{k+1/2}(\theta))\right) + \tau\left(F(\mathbf{z}_{k+1/2}(\theta)) - F(\mathbf{w}_k(\theta))\right)&\\
\in \tau\left(F+G\right)(\mathbf{z}_{k+1/2}(\theta)),&
\end{aligned} \tag{52}$$

**Algorithm 4** FoRB with variance reduction

**Input:** Probability $p \in (0,1]$, probability distribution $Q$, step size $\tau$, $\alpha \in (0,1)$. Let $\mathbf{z}_0 = \mathbf{w}_0$

**for** $k = 1, 2 \dots$ **do**

$\bar{\mathbf{z}}_k = \alpha \mathbf{z}_k + (1-\alpha)\mathbf{w}_k$

Draw an index $\xi_k$ according to $Q$

$\mathbf{z}_{k+1} = J_{\tau G}(\bar{\mathbf{z}}_k - \tau F(\mathbf{w}_k) - \tau(F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1})))$

$$\mathbf{w}_{k+1} = \begin{cases} \mathbf{z}_{k+1}, & \text{with probability } p \\ \mathbf{w}_k, & \text{with probability } 1-p \end{cases}$$

**end for**

to show that $\tilde{\mathbf{z}}(\theta) \in (F+G)^{-1}(0)$. In particular, we use that $F_\xi$ is continuous for all $\xi$, $\mathbf{z}_{k+1} - \bar{\mathbf{z}}_k \to 0$, and $\mathbf{z}_{k+1/2} - \mathbf{w}_k \to 0$ almost surely. We use the same arguments as the proof of Theorem 2 to conclude. ■

We next give the complexity of the algorithm for solving VI as Section 2.3.1. The derivation is essentially the same as Section 2.3.1 and therefore omitted.

**Corollary 20** *Let* $\alpha = 1 - p = 1 - \frac{2}{N}$ *and* $\mathbf{z}^K = \frac{1}{K}\sum_{k=0}^{K-1} \mathbf{z}_{k+1/2}$. *Then, the total complexity to get an* $\varepsilon$*-accurate solution to* (1) *is* $\mathcal{O}\left(N + \frac{\sqrt{N}L}{\epsilon}\right)$.

### C.2. Forward-reflected-backward with variance reduction: revisited

In a similar spirit to FBF, but using a different idea, Malitsky and Tam (2020) proposed FoRB method

$$\mathbf{z}_{k+1} = J_{\tau G}\left(\mathbf{z}_k - \tau[F(\mathbf{z}_k) + F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})]\right).$$

This scheme generalizes optimistic gradient descent (Rakhlin and Sridharan, 2013; Daskalakis et al., 2018) and in some particular cases is equivalent to Popov's method (Popov, 1980). Later, in (Alacaoglu et al., 2021), the authors suggested the most straightforward variance reduction modification of FoRB by combining FoRB and loopless SVRG (Kovalev et al., 2020). This algorithm had the drawback of small step sizes which lead to complexity bounds that do not improve upon the deterministic methods. As highlighted in the experiments of (Alacaoglu et al., 2021), the small step size $\tau \sim \frac{1}{n}$ seemed to be non-improvable for the given method. One possible speculation for this phenomenon might be that the method is too aggressive and therefore prohibits large step sizes. We will use the retracted iterate $\bar{\mathbf{z}}_k = \alpha \mathbf{z}_k + (1-\alpha)\mathbf{w}_k$ instead of the latest iterate $\mathbf{z}_k$ in the update to improve complexity.

The advantage of FoRB compared to extragradient is similar to FBF. FoRB only needs one proximal operator, applied to VI. Compared to FBF, FoRB has a simpler update rule and, unlike FBF, it is easy to adjust to Bregman setting, see (Alacaoglu et al., 2021; Zhang, 2022).

**Remark 21** *For running Alg. 4 in practice, we suggest* $p = \frac{2}{N}$, $\alpha = 1 - p$, *and* $\tau = \frac{0.99\sqrt{p(1-p)}}{L}$.

Lyapunov function here is slightly more complicated than the ones in previous sections:

$$\Phi_{k+1}(\mathbf{z}) \coloneqq \alpha\|\mathbf{z}_{k+1}-\mathbf{z}\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_{k+1}-\mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_{k+1})-F(\mathbf{w}_k), \mathbf{z}-\mathbf{z}_{k+1}\rangle \\ + (1-\alpha)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2.$$

**Theorem 22** *Let Assumption 3 hold, $\alpha \in [0,1)$, $p \in (0,1]$, and $\tau = \frac{\sqrt{\alpha(1-\alpha)}}{L}\gamma$ for $\gamma \in (0,1)$. Then for $(\mathbf{z}_k)$ generated by Alg. 4 and any $\mathbf{z}_* \in \mathsf{Sol}$, it holds that $\Phi_k(\mathbf{z}_*)$ is nonnegative and*

$$\mathbb{E}_k\left[\Phi_{k+1}(\mathbf{z}_*)\right] \le \Phi_k(\mathbf{z}_*).$$

*Moreover, if $F_\xi$ is continuous for all $\xi$, then $(\mathbf{z}_k)$ converges to some $\mathbf{z}_* \in \mathsf{Sol}$ a.s.*

**Remark 23** *Note that again when randomness is null, $F_\xi = F$ and $p = 1$, Alg. 4 reduces to the original FoRB algorithm. Moreover, with $\alpha = \frac{1}{2}$ we recover the result in (Malitsky and Tam, 2020).*

**Proof of Theorem 22** Nonnegativity of $\Phi_k(\mathbf{z}_*)$ is straightforward to prove by using Lipschitzness of $F$ and $\tau L \le \sqrt{a(1-\alpha)}$.

Let $\mathbf{z} = \mathbf{z}_* \in \mathsf{Sol}$ which gives $-F(\mathbf{z}) \in G(\mathbf{z})$. Next, by the definitions of $\mathbf{z}_{k+1}$ and resolvent, $\bar{\mathbf{z}}_k - \tau\left[F(\mathbf{w}_k) - F_{\xi_k}(\mathbf{w}_{k-1}) + F_{\xi_k}(\mathbf{z}_k)\right] \in \mathbf{z}_{k+1} + \tau G(\mathbf{z}_{k+1})$. Combining these estimates and monotonicity of $G$ leads to

$$\langle \mathbf{z}_{k+1} - \bar{\mathbf{z}}_k + \tau\left[F(\mathbf{w}_k) - F_{\xi_k}(\mathbf{w}_{k-1}) + F_{\xi_k}(\mathbf{z}_k)\right], \mathbf{z}-\mathbf{z}_{k+1}\rangle - \tau\langle F(\mathbf{z}), \mathbf{z}-\mathbf{z}_{k+1}\rangle \ge 0. \quad (53)$$

We split the first inner product and work with each term separately. First,

$$\begin{aligned}
&\tau\langle F(\mathbf{w}_k) - F_{\xi_k}(\mathbf{w}_{k-1}) + F_{\xi_k}(\mathbf{z}_k), \mathbf{z}-\mathbf{z}_{k+1}\rangle \\
&\quad= \tau\langle F(\mathbf{w}_k) - F(\mathbf{z}_{k+1}), \mathbf{z}-\mathbf{z}_{k+1}\rangle - \tau\langle F_{\xi_k}(\mathbf{w}_{k-1}) - F_{\xi_k}(\mathbf{z}_k), \mathbf{z}-\mathbf{z}_{k+1}\rangle \\
&\qquad+ \tau\langle F(\mathbf{z}_{k+1}), \mathbf{z}-\mathbf{z}_{k+1}\rangle \\
&\quad= \tau\langle F(\mathbf{w}_k) - F(\mathbf{z}_{k+1}), \mathbf{z}-\mathbf{z}_{k+1}\rangle - \tau\langle F_{\xi_k}(\mathbf{w}_{k-1}) - F_{\xi_k}(\mathbf{z}_k), \mathbf{z}-\mathbf{z}_k\rangle \\
&\qquad- \tau\langle F_{\xi_k}(\mathbf{w}_{k-1}) - F_{\xi_k}(\mathbf{z}_k), \mathbf{z}_k-\mathbf{z}_{k+1}\rangle + \tau\langle F(\mathbf{z}_{k+1}), \mathbf{z}-\mathbf{z}_{k+1}\rangle.
\end{aligned}$$

Second, as we derived in (6),

$$2\langle \mathbf{z}_{k+1}-\bar{\mathbf{z}}_k, \mathbf{z}-\mathbf{z}_{k+1}\rangle = \alpha\|\mathbf{z}_k-\mathbf{z}\|^2 - \|\mathbf{z}_{k+1}-\mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k-\mathbf{z}\|^2 \\ - \alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2 - (1-\alpha)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2.$$

Substituting the last two estimates into (53), we obtain

$$\begin{aligned}
&\|\mathbf{z}_{k+1}-\mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_{k+1}) - F(\mathbf{w}_k), \mathbf{z}-\mathbf{z}_{k+1}\rangle + 2\tau\langle F(\mathbf{z}) - F(\mathbf{z}_{k+1}), \mathbf{z}-\mathbf{z}_{k+1}\rangle \\
&\le \alpha\|\mathbf{z}_k-\mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k-\mathbf{z}\|^2 + 2\tau\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}-\mathbf{z}_k\rangle \\
&\quad+ 2\tau\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}_k-\mathbf{z}_{k+1}\rangle - \alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2 - (1-\alpha)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2. \quad (54)
\end{aligned}$$

We take expectation conditioning on the knowledge of $\mathbf{z}_k, \mathbf{w}_k$, use $\mathbb{E}_k F_{\xi_k}(\mathbf{z}_k) = F(\mathbf{z}_k), \mathbb{E}_k F_{\xi_k}(\mathbf{w}_{k-1}) = F(\mathbf{w}_{k-1})$, and monotonicity of $F$ for the third term in the LHS. This yields

$$\begin{aligned}
&\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_{k+1}) - F(\mathbf{w}_k), \mathbf{z}-\mathbf{z}_{k+1}\rangle + (1-\alpha)\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2\right] \\
&\le \alpha\|\mathbf{z}_k-\mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k-\mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_k) - F(\mathbf{w}_{k-1}), \mathbf{z}-\mathbf{z}_k\rangle \\
&\quad+ 2\tau\,\mathbb{E}_k\left[\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}_k-\mathbf{z}_{k+1}\rangle - \alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2\right]. \quad (55)
\end{aligned}$$

Using Assumption 1(iv), Cauchy-Schwarz and Young's inequalities, we can bound the last line above as

$$\begin{aligned}
&\mathbb{E}_k[2\tau\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}_k - \mathbf{z}_{k+1}\rangle - \alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2] \\
&\leq \mathbb{E}_k\left[\frac{\tau^2}{\alpha\gamma}\|F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1})\|^2 + \alpha\gamma\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 - \alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2\right] \\
&\leq \frac{(1-\alpha)\gamma}{L^2}\,\mathbb{E}_k\,\|F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1})\|^2 - (1-\gamma)\alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\
&\leq (1-\alpha)\gamma\|\mathbf{z}_k - \mathbf{w}_{k-1}\|^2 - (1-\gamma)\alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2. \qquad (56)
\end{aligned}$$

Adding (11) and (56) to (55), we obtain

$$\mathbb{E}_k[\Phi_{k+1}(\mathbf{z})] \leq \Phi_k(\mathbf{z}) - (1-\alpha)(1-\gamma)\|\mathbf{z}_k - \mathbf{w}_{k-1}\|^2 - (1-\gamma)\alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2.$$

The rest of the proof is the same as Theorem 19. The only difference is that instead of (52), we have

$$\begin{aligned}
\mathbf{z}_{k+1}(\theta) - \bar{\mathbf{z}}_k(\theta) + \tau\left(F_{\xi_k}(\mathbf{z}_k(\theta)) - F_{\xi_k}(\mathbf{w}_{k-1}(\theta))\right) + \tau\left(F(\mathbf{z}_{k+1}(\theta)) - F(\mathbf{w}_k(\theta))\right) \\
\in \tau\left(F+G\right)(\mathbf{z}_{k+1}(\theta)),
\end{aligned}$$

which gives the same conclusion as $F_\xi$ is continuous for all $\xi$, $\mathbf{z}_{k+1} - \bar{\mathbf{z}}_k \to 0$, $\mathbf{z}_{k+1} - \mathbf{w}_k \to 0$ almost surely. ■

**Remark 24** *Even though we will set the parameters $\alpha$, $p$, $\tau$ by optimizing complexity, we observe that the requirements in Theorem 22 allows step sizes arbitrary close to $\frac{1}{2L}$. This already shows flexibility of the analysis, compared to the strict requirement of $\tau = \frac{p}{4L}$ in (Alacaoglu et al., 2021).*

The improvement in the step size choice is due to using $\bar{\mathbf{z}}_k$ which allows us to use tighter estimations whereas the analysis in (Alacaoglu et al., 2021) needs to make use of multiple Young's inequalities. In particular, we use $\mathbf{z}_*$ as an anchor point in (11), whereas (Alacaoglu et al., 2021) uses $\mathbf{z}_k$ as anchor point, which requires Young's inequalities to transform to $\mathbf{z}_{k-1}$ and obtain a telescoping sum. Finally, as Corollary 20, we give the complexity of the algorithm for solving VI in the spirit of Section 2.3.1.

**Corollary 25** *Let $\alpha = 1 - p = 1 - \frac{2}{N}$ and $\mathbf{z}^K = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{z}_k$. Then, the total complexity to get an $\varepsilon$-accurate solution to (1) is $\mathcal{O}\left(N + \frac{\sqrt{N}L}{\epsilon}\right)$.*

### C.3. Linear convergence

In this section, we illustrate how to obtain linear convergence of Alg. 1 for solving VI (1) when $g$ is $\mu$-strongly convex. Alternatively, one can replace this assumption with strong monotonicity of $F$, which we omit for brevity. One can use the same arguments for FBF and FoRB variants in the previous sections to show linear convergence for solving strongly monotone inclusions.

**Theorem 26** *Let Assumption 1 hold, $g$ be $\mu$-strongly convex, and $\mathbf{z}_*$ be the solution of (1). If we set $\alpha = 1 - p$ and $\tau = \frac{\sqrt{p}}{2L}$ in Alg. 1, then it holds that*

$$\mathbb{E}\|\mathbf{z}_k - \mathbf{z}_*\|^2 \leq \left(\frac{1}{1 + c/3}\right)^k \frac{2}{1-p}\|\mathbf{z}_0 - \mathbf{z}_*\|^2,$$

*with $c = \min\left\{\frac{3p}{8}, \frac{\sqrt{p}\mu}{2L}\right\}$.*

**Proof** In (4), we use strong convexity of $g$ to have an additional term $\frac{\tau\mu}{2}\|\mathbf{z}_{k+1}-\mathbf{z}\|^2$ on the right-hand side of the first inequality. Next, we continue as in the proof of Lemma 1 to obtain, instead of (10),

$$(1+\tau\mu)\,\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2\right] \leq \alpha\|\mathbf{z}_k-\mathbf{z}_*\|^2+(1-\alpha)\|\mathbf{w}_k-\mathbf{z}_*\|^2-(1-\alpha)(1-\gamma)\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2 \\ -(1-\gamma)\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right].$$

We add (11) to this inequality after using the tower property, to deduce

$$(\alpha+\tau\mu)\,\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2\right]+\frac{1-\alpha}{p}\mathbb{E}_k\left[\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2\right] \leq \alpha\|\mathbf{z}_k-\mathbf{z}_*\|^2+\frac{1-\alpha}{p}\|\mathbf{w}_k-\mathbf{z}_*\|^2 \\ -(1-\gamma)\Big((1-\alpha)\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right]\Big).$$

Since we set $\alpha=1-p$ and $\gamma=\frac{1}{2}$, we can rewrite it as

$$(1-p+\tau\mu)\,\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2\right]+\mathbb{E}_k\left[\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2\right] \leq (1-p)\|\mathbf{z}_k-\mathbf{z}_*\|^2+\|\mathbf{w}_k-\mathbf{z}_*\|^2 \\ -\frac{1}{2}\left(p\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right]\right). \quad (57)$$

Next, by $2\|u\|^2+2\|v\|^2\geq\|u+v\|^2$ applied two times,

$$\begin{aligned}\frac{2c}{3}\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2 &\geq \frac{c}{3}\mathbb{E}_k\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2-\frac{2c}{3}\mathbb{E}_k\left[\mathbb{E}_{k+1/2}\|\mathbf{z}_{k+1}-\mathbf{w}_{k+1}\|^2\right]\\ &=\frac{c}{3}\mathbb{E}_k\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2-\frac{2c(1-p)}{3}\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2\\ &\geq\frac{c}{3}\mathbb{E}_k\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2-\frac{4c}{3}\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2-\frac{4c}{3}\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2.\end{aligned}$$

Using this inequality in (57) and that $c\leq\frac{\sqrt{p}\mu}{2L}=\tau\mu$ gives us

$$\left(1-p+\frac{c}{3}\right)\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2\right]+\left(1+\frac{c}{3}\right)\mathbb{E}_k\left[\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2\right] \leq (1-p)\|\mathbf{z}_k-\mathbf{z}_*\|^2+\|\mathbf{w}_k-\mathbf{z}_*\|^2 \\ -\frac{1}{2}\left(p\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right)+\frac{4c}{3}\left(\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right). \quad (58)$$

By our choice of $c$, we have $\frac{4c}{3}\leq\frac{p}{2}$ and, therefore, the second line of (58) is nonpositive. Using $1-p+\frac{c}{3}>(1-p)(1+\frac{c}{3})$ and taking total expectation, yields

$$\left(1+\frac{c}{3}\right)\mathbb{E}\left[(1-p)\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2+\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2\right] \leq \mathbb{E}\left[(1-p)\|\mathbf{z}_k-\mathbf{z}_*\|^2+\|\mathbf{w}_k-\mathbf{z}_*\|^2\right].$$

By iterating this inequality, we obtain

$$(1-p)\mathbb{E}\|\mathbf{z}_k-\mathbf{z}_*\|^2\leq\left(\frac{1}{1+c/3}\right)^k(2-p)\|\mathbf{z}_0-\mathbf{z}_*\|^2,$$

which gives the result. ■

**Corollary 27** *Let $p = \frac{2}{N}$, $\tau = \frac{\sqrt{p}}{2L}$. The total average complexity is $\mathcal{O}\left(\left(N + \frac{\sqrt{NL}}{\mu}\right)\log\frac{1}{\epsilon}\right)$.*

**Proof** The $\varepsilon$-accuracy is reached after $\mathcal{O}(\log\frac{1}{\varepsilon}/\log(1+\frac{c}{3}))$ iterations. This yields a factor $\frac{pN+2}{\log(1+\frac{c}{3})} \approx \frac{3}{c}(pN+2)$ in total complexity. Using our choice for $c$, we obtain total average complexity

$$\max\left\{\frac{8}{p}, \frac{6L}{\sqrt{p}\mu}\right\}(pN+2) \leq \frac{32}{p} + \frac{24L}{\sqrt{p}\mu} = 16N + \frac{12\sqrt{2N}L}{\mu}.$$

We lastly multiply the last estimate with $\log\left(\varepsilon^{-1}\right)$. ■

**Remark 28** *In this case, Alg. 1 has complexity $\mathcal{O}\left(\left(N + \frac{\sqrt{NL}}{\mu}\right)\log\frac{1}{\varepsilon}\right)$, compared to the deterministic methods $\mathcal{O}\left(\frac{NL_F}{\mu}\log\frac{1}{\varepsilon}\right)$. This complexity recovers the previously obtained result in (Balamurugan and Bach, 2016) and (Carmon et al., 2019, Section 5.4), where our advantage is having algorithmic parameters independent of $\mu$ and having more general assumptions.*

## Appendix D. Applications

### D.1. Bilinear min-max problems

In this section, we analyze the overall complexity of our method compared to deterministic extragradient and show the complexity improvements.

**Notation.** For a vector $\mathbf{x}$ we use $x_i$ to denote its $i$-th coordinate and for an indexed vector $\mathbf{x}_k$ it is $x_{k,i}$. For a matrix $A \in \mathbb{R}^{m\times n}$ we denote a number of its non-zero entries by $\mathrm{nnz}(A)$; it is exactly the complexity of computing $A\mathbf{x}$ or $A^\top\mathbf{y}$. We use the spectral, Frobenius and max norms of $A$ defined as $\|A\| = \sigma_{\max}(A)$, $\|A\|_{\mathrm{Frob}} = \sqrt{\sum_{i,j}A_{ij}^2} = \sqrt{\sum_{i=1}^{\mathrm{rank}(A)}\sigma_i(A)^2}$, and $\|A\|_{\max} = \max_{i,j}|A_{ij}|$. For $i$-th row and $j$-th column of $A$ we use a convenient notation $A_{i:}$ and $A_{:j}$. Here, for simplicity, we measure complexity in terms of arithmetic operations.

**Problem.** The general problem that we consider is

$$\min_{\mathbf{x}\in\mathbb{R}^n}\max_{\mathbf{y}\in\mathbb{R}^m}\langle A\mathbf{x},\mathbf{y}\rangle + g_1(\mathbf{x}) - g_2(\mathbf{y}),$$

where $g_1, g_2$ are proper convex lsc functions. We can formulate this problem as a VI by setting

$$F(\mathbf{z}) = F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} A^\top\mathbf{y} \\ -A\mathbf{x}\end{pmatrix}, \qquad g(\mathbf{z}) = g_1(\mathbf{x}) + g_2(\mathbf{y}). \tag{59}$$

#### D.1.1. LINEARLY CONSTRAINED MINIMIZATION

A classical example of bilinear saddle point problems is linearly constrained minimization

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) : A\mathbf{x} = b,$$

where $f$ is proper convex lsc. The equivalent min-max formulation corresponds to (59) when $g_1(\mathbf{x}) = f(\mathbf{x})$ and $g_2(\mathbf{y}) = \langle b,\mathbf{y}\rangle$.

We will instantiate Alg. 1 for this problem. To make our presentation clearer, we consider only the most common scenario when $\operatorname{nnz}(A) > m + n$. In this setting, deterministic methods (extragradient, FBF, FoRB, etc.) solve (62) with $\mathcal{O}\left(\operatorname{nnz}(A)\|A\|\varepsilon^{-1}\right)$ total complexity. As we see in the sequel, variance reduced methods provide us $\mathcal{O}\left(\operatorname{nnz}(A) + \sqrt{\operatorname{nnz}(A)(m+n)}\|A\|_{\text{Frob}}\varepsilon^{-1}\right)$ total complexity. We now describe the definition of $F_\xi$ with two oracle choices. The first choice is the version of "importance" sampling described in Section 2.1.

**Oracle 1.** The fixed distribution (the same in every iteration) is defined as

$$F_\xi(\mathbf{z}) = \begin{pmatrix} \frac{1}{r_i} A_{i:} y_i \\ -\frac{1}{c_j} A_{:j} x_j \end{pmatrix}, \qquad \Pr\{\xi = (i,j)\} = r_i c_j, \quad r_i = \frac{\|A_{i:}\|_2^2}{\|A\|_{\text{Frob}}^2}, \quad c_j = \frac{\|A_{:j}\|_2^2}{\|A\|_{\text{Frob}}^2}.$$

In the view of Assumption 1, the Lipschitz constant of $F_\xi$ can be computed as

$$\begin{aligned} \mathbb{E}\,\|F_\xi(\mathbf{z})\|_2^2 &= \mathop{\mathbb{E}}_{i\sim r}\left[\frac{1}{r_i^2}\|A_{i:}y_i\|_2^2\right] + \mathop{\mathbb{E}}_{j\sim c}\left[\frac{1}{c_j^2}\|A_{:j}x_j\|_2^2\right] = \sum_{i=1}^m \frac{1}{r_i}\|A_{i:}y_i\|_2^2 + \sum_{j=1}^n \frac{1}{c_j}\|A_{:j}x_j\|_2^2 \\ &= \sum_{i=1}^m \frac{1}{r_i}\|A_{i:}\|_2^2 (y_i)^2 + \sum_{j=1}^n \frac{1}{c_j}\|A_{:j}\|_2^2 (x_j)^2 = \|A\|_{\text{Frob}}^2\|\mathbf{z}\|_2^2. \end{aligned} \tag{60}$$

**Oracle 2.** The second stochastic oracle is slightly more complicated, since it is iteration-dependent as (Carmon et al., 2019). We use the setting of Assumption 2. Given $\mathbf{u} = (\mathbf{u}^x, \mathbf{u}^y)$ and $\mathbf{v} = (\mathbf{v}^x, \mathbf{v}^y)$, for $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, we define

$$F_\xi(\mathbf{z}) = \begin{pmatrix} \frac{1}{r_i} A_{i:} y_i \\ -\frac{1}{c_j} A_{:j} x_j \end{pmatrix}, \qquad \Pr\{\xi = (i,j)\} = r_i c_j, \quad r_i = \frac{|u_i^y - v_i^y|^2}{\|\mathbf{u}^y - \mathbf{v}^y\|^2}, \quad c_j = \frac{|u_j^x - v_j^x|^2}{\|\mathbf{u}^x - \mathbf{v}^x\|^2},$$

and call the described distribution as $Q(\mathbf{u}, \mathbf{v})$. Similarly, in every iteration of Alg. 2 we define a distribution $Q(\mathbf{z}_{k+1/2}^s, \mathbf{w}^s)$ and sample $\xi$ according to it.

Clearly, as before, $F_\xi$ is unbiased. It is easy to show that this oracle is variable $\|A\|_{\text{Frob}}$-Lipschitz. Its proof is similar to the variable Lipschitz derivation that we will include for matrix games with Bregman distances, in Section D.1.2.

**Complexity.** We suppose that computing proximal operators $\operatorname{prox}_{g_1}$, $\operatorname{prox}_{g_2}$ can be done efficiently in $\tilde{\mathcal{O}}(m+n)$ complexity. Our result in Theorem 4 stated that Alg. 1 has the rate $\mathcal{O}\left(\frac{L}{\sqrt{p}K}\right)$. Given that the expected cost of each iteration is $\mathcal{O}\left(p\operatorname{nnz}(A) + m + n\right)$, setting $p = \frac{m+n}{\operatorname{nnz}(A)}$ gives us the average total complexity

$$\tilde{\mathcal{O}}\left(\operatorname{nnz}(A) + \frac{\sqrt{\operatorname{nnz}(A)(m+n)}\|A\|_{\text{Frob}}}{\varepsilon}\right). \tag{61}$$

It is easy to see that Alg. 2 has the same complexity if we set $K = \left\lceil \frac{\operatorname{nnz}(A)}{m+n} \right\rceil$.

Compared to the complexity of deterministic methods, this complexity improves depending on the relation between $\|A\|_{\text{Frob}}$ and $\|A\|$. In particular, when $A$ is a square dense matrix, due to $\|A\|_{\text{Frob}} \leq \sqrt{\operatorname{rank}(A)}\|A\|$, the bound in (61) improves that of deterministic method. In (61) we suppress $\|\mathbf{z}_0 - \mathbf{z}_*\|^2$ that is common to both our methods and deterministic ones.

Finally, we remark that the analysis in (Carmon et al., 2019, Section 5.2) requires the additional assumption that $\mathbf{z} \mapsto \langle F(\mathbf{z}) + \tilde{\nabla} f(\mathbf{z}), \mathbf{z} - \mathbf{u}\rangle$ is convex for all $\mathbf{u}$ to apply to this case, where we denote a subgradient of $f$ by $\tilde{\nabla} f$. This assumption requires more structure on $f$.

### D.1.2. MATRIX GAMES

The problem in this case is written as

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \langle A\mathbf{x}, \mathbf{y}\rangle, \tag{62}$$

where $A \in \mathbb{R}^{m\times n}$ and $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$ are closed convex sets, projection onto each are easy to compute. In view of (59), we have $g(\mathbf{z}) = \delta_{\mathcal{X}}(\mathbf{x}) + \delta_{\mathcal{Y}}(\mathbf{y})$. As we shall see, our complexities in this case recover the ones in (Carmon et al., 2019). We refer to Section 1.1 for a detailed comparison.

In the Euclidean setup, we suppose that the underlying space $\mathcal{Z} = \mathbb{R}^n \times \mathbb{R}^m$ has a Euclidean structure with the norm $\|\cdot\|_2$ and, hence, it coincides with the dual $\mathcal{Z}^*$. In this case, we can use Oracle 1 and Oracle 2 from Section D.1.1 and we obtain the same complexity as (61). The same discussions as Section D.1.1 apply.

#### BREGMAN SETUP

Let $\mathcal{X} = \Delta^n = \{\mathbf{x} \in \mathbb{R}^n \colon \sum_{i=1}^n x_i = 1, x_i \geq 0\}$ and $\mathcal{Y} = \Delta^m$. With this, problem (62) is known as a zero sum game. In this case, deterministic algorithms formulated with a specific Bregman distance (given below) have $\mathcal{O}\left(\mathrm{nnz}(A)\|A\|_{\max}\varepsilon^{-1}\right)$ total complexity. These settings are standard and we recall them only for reader's convenience.

For $\mathcal{Z} = \mathbb{R}^{m+n}$ and $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ we define $\|\mathbf{z}\| = \sqrt{\|\mathbf{x}\|_1^2 + \|\mathbf{y}\|_1^2}$. Correspondingly, $\mathcal{Z}^* = (\mathbb{R}^{m+n}, \|\cdot\|_*)$ is the dual space with $\|\mathbf{z}^*\| = \sqrt{\|\mathbf{x}^*\|_\infty^2 + \|\mathbf{y}^*\|_\infty^2}$ for $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$. For $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \Delta^n \times \Delta^m$ we use the negative entropy $h_1(\mathbf{x}) = \sum_{i=1}^n x_i \log x_i$, $h_2(\mathbf{y}) = \sum_{i=1}^m y_i \log y_i$ and set $h(\mathbf{z}) = h_1(\mathbf{x}) + h_2(\mathbf{y}) = \sum_{i=1}^{m+n} z_i \log z_i$. Then we define the Bregman distance as

$$D(\mathbf{z}, \mathbf{z}') = h(\mathbf{z}) - h(\mathbf{z}') - \langle \nabla h(\mathbf{z}'), \mathbf{z} - \mathbf{z}'\rangle = \sum_i z_i \log \frac{z_i}{z_i'}.$$

Of course, this definition requires $\mathbf{z}'$ to be in the relative interior of $\Delta^n \times \Delta^m$; normally it is satisfied automatically for the iterates of the algorithm (including our Alg. 2).

If we choose $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_0)$ with $\mathbf{x}_0 = \frac{1}{n}\mathbb{1}_n$, $\mathbf{y}_0 = \frac{1}{m}\mathbb{1}_m$, it is easy to see that

$$\max_{\mathbf{z}\in\Delta^n\times\Delta^m} D(\mathbf{z}, \mathbf{z}_0) \leq \log n + \log m = \log(mn).$$

We know that $D$ satisfies $D(\mathbf{z}, \mathbf{z}') \geq \frac{1}{2}\|\mathbf{z} - \mathbf{z}'\|^2$ for all $\mathbf{z}, \mathbf{z}' \in \Delta^n \times \Delta^m$. Deterministic algorithms have constant $\|A\|_{\max}$ in their complexity, since $F$ defined in (59) is $\|A\|_{\max}$-Lipschitz:

$$\|F(\mathbf{z})\|_*^2 = \|A^\top \mathbf{y}\|_\infty^2 + \|A\mathbf{x}\|_\infty^2 \leq \|A\|_{\max}^2(\|\mathbf{x}\|_1^2 + \|\mathbf{y}\|_1^2) = \|A\|_{\max}^2\|\mathbf{z}\|^2.$$

**Oracle.** The stochastic oracle here is similar to the Oracle 2 in Section D.1.1 for the Euclidean case, but with adjustment to the $\ell_1$-norm. Again we are in the setting of Assumption 2. Given $\mathbf{u} = (\mathbf{u}^x, \mathbf{u}^y)$ and $\mathbf{v} = (\mathbf{v}^x, \mathbf{v}^y)$, for $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, we define

$$F_\xi(\mathbf{z}) = \begin{pmatrix} \frac{1}{r_i} A_{i:} y_i \\ -\frac{1}{c_j} A_{:j} x_j \end{pmatrix}, \qquad \Pr\{\xi = (i,j)\} = r_i c_j, \quad r_i = \frac{|u_i^y - v_i^y|}{\|\mathbf{u}^y - \mathbf{v}^y\|_1}, \quad c_j = \frac{|u_j^x - v_j^x|}{\|\mathbf{u}^x - \mathbf{v}^x\|_1},$$

and call the described distribution as $Q(\mathbf{u},\mathbf{v})$. We show that $F_\xi$ is variable $\|A\|_{\max}$-Lipschitz in view of Definition 7. Indeed, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\xi\sim Q(\mathbf{u},\mathbf{v})}[\|F_\xi(\mathbf{u})-F_\xi(\mathbf{v})\|_*^2] &= \mathop{\mathbb{E}}_{\xi\sim Q(\mathbf{u},\mathbf{v})}\left[\|F_\xi(\mathbf{u}-\mathbf{v})\|_*^2\right] \\
&= \mathop{\mathbb{E}}_{i\sim r}\left[\frac{1}{r_i^2}\|A_{i:}(u_i^y-v_i^y)\|_{\max}^2\right] + \mathop{\mathbb{E}}_{j\sim c}\left[\frac{1}{c_j^2}\|A_{:j}(u_j^x-v_j^x)\|_{\max}^2\right] \\
&= \sum_{i=1}^m \frac{1}{r_i}\|A_{i:}\|_{\max}^2|u_i^y-v_i^y|^2 + \sum_{j=1}^n \frac{1}{c_j}\|A_{:j}\|_{\max}^2|u_j^x-v_j^x|^2 \\
&\le \sum_{i=1}^m \|A\|_{\max}^2|u_i^y-v_i^y|\|\mathbf{u}^y-\mathbf{v}^y\|_1 + \sum_{j=1}^n \|A\|_{\max}^2|u_j^x-v_j^x|\|\mathbf{u}^x-\mathbf{v}^x\|_1 \\
&= \|A\|_{\max}^2\left(\|\mathbf{u}^y-\mathbf{v}^y\|_1^2+\|\mathbf{u}^y-\mathbf{v}^x\|_1^2\right) = \|A\|_{\max}^2\|\mathbf{u}-\mathbf{v}\|^2.
\end{aligned}$$

Similarly, in every iteration of Alg. 2 we define a distribution $Q(\mathbf{z}_{k+1/2}^s,\mathbf{w}^s)$ and sample $\xi_k^s$ according to it. This stochastic oracle was already used in (Grigoriadis and Khachiyan, 1995) and used extensively after that, see (Nesterov and Nemirovski, 2013; Clarkson et al., 2012) and references therein. In (Carmon et al., 2019) this oracle was called "sampling from the difference".

**Complexity.** In this case, the complexity of deterministic algorithms (Mirror Prox, FoRB) is $\mathcal{O}\left(\mathrm{nnz}(A)\|A\|_{\max}\varepsilon^{-1}\right)$. Our result in Corollary 9 stated that Alg. 2 has the rate $\mathcal{O}\left(\frac{L}{\sqrt{KS}}\right)$. Given that the cost of each epoch of Alg. 2 is $\mathcal{O}\left(\mathrm{nnz}(A)+K(m+n)\right)$, setting $K=\left\lceil\frac{\mathrm{nnz}(A)}{m+n}\right\rceil$ gives us the total complexity

$$\tilde{\mathcal{O}}\left(\mathrm{nnz}(A)+\frac{\sqrt{\mathrm{nnz}(A)(m+n)}\|A\|_{\max}}{\varepsilon}\right),$$

which, in the square dense case, improves the deterministic complexity by $\sqrt{n}$.

**Updates.** For concreteness we specify updates in lines 4–7 of Alg. 2. Let $\mathbf{w}^s=(\mathbf{u},\mathbf{v})$, $\bar{\mathbf{w}}^s=(\bar{\mathbf{u}}^s,\bar{\mathbf{v}}^s)$.

$$\begin{aligned}
\nabla h_1(\mathbf{x}_{k+1/2}^s) &= \alpha\nabla h_1(\mathbf{x}_k^s)+(1-\alpha)\nabla h_1(\bar{\mathbf{u}}^s)-\tau A^\top\mathbf{v}^s \\
\nabla h_2(\mathbf{y}_{k+1/2}^s) &= \alpha\nabla h_2(\mathbf{y}_k^s)+(1-\alpha)\nabla h_2(\bar{\mathbf{v}}^s)+\tau A\mathbf{u}^s
\end{aligned}$$

Then we form a distribution $Q(\mathbf{z}_{k+1/2}^s,\mathbf{w}^s)$

$$\Pr\{\xi=(i,j)\}=r_ic_j,\quad r_i=\frac{|y_{k+1/2,i}^s-v_i^s|}{\|\mathbf{y}_{k+1/2}^s-\mathbf{v}^s\|_1},\quad c_j=\frac{|x_{k+1/2,i}^s-u_i^s|}{\|\mathbf{x}_{k+1/2}^s-\mathbf{u}^s\|_1}$$

and sample $\xi_k=(i,j)$ according to $Q(\mathbf{z}_{k+1/2}^s,\mathbf{w}^s)$. Finally, we update $\mathbf{x}_{k+1}^s$ and $\mathbf{y}_{k+1}^s$ as

$$\begin{aligned}
\nabla h_1(\mathbf{x}_{k+1}^s) &= \alpha\nabla h_1(\mathbf{x}_k^s)+(1-\alpha)\nabla h_1(\bar{\mathbf{u}}^s)-\tau A^\top\mathbf{v}^s-\frac{\tau}{r_i}A_{i:}(y_{k+1/2,i}^s-v_i^s) \\
&= \nabla h_1(\mathbf{x}_{k+1/2}^s)-\tau A_{i:}\|\mathbf{y}_{k+1/2}^s-\mathbf{v}^s\|\,\mathrm{sign}(y_{k+1/2,i}^s-v_i^s) \\
\nabla h_2(\mathbf{y}_{k+1}^s) &= \nabla h_2(\mathbf{y}_{k+1/2}^s)+\tau A_{:j}\|\mathbf{x}_{k+1/2}^s-\mathbf{u}^s\|\,\mathrm{sign}(x_{k+1/2,j}^s-u_j^s)
\end{aligned}$$

Switching from dual variables $\nabla h_1(\mathbf{x})$ to primal $\mathbf{x}$ is elementary by duality:

$$X = \nabla h_1(\mathbf{x}) \quad \iff \quad \mathbf{x} = \nabla h_1^*(X) = \frac{(e^{X_1}, \dots, e^{X_n})}{\sum_{i=1}^n e^{X_i}}$$

and similarly for $\mathbf{y}$. Updates for $\mathbf{w}$ and $\nabla h(\bar{\mathbf{w}})$ are straightforward by means of incremental averaging.

### D.2. Nonbilinear min-max problems

An important example of nonbilinear min-max problems is constrained optimization

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \quad \text{subject to} \quad h_i(\mathbf{x}) \le 0, \text{ for } i \in [N],$$

where $f, h_i$ are smooth convex functions. We can map this problem to the VI template (1) by setting

$$F = \begin{pmatrix} \nabla f(\mathbf{x}) + \sum_{i=1}^N y_i \nabla h_i(\mathbf{x}) \\ -\big(h_1(\mathbf{x}), \dots h_N(\mathbf{x})\big)^\top \end{pmatrix}, \qquad g(\mathbf{z}) = \delta_{\mathcal{X}}(\mathbf{x}) + \delta_{\mathbb{R}^N_+}(\mathbf{y}).$$

One possible choice for stochastic oracles is to set

$$F_i(\mathbf{z}) = \begin{pmatrix} \nabla f(\mathbf{x}) + N y_i \nabla h_i(\mathbf{x}) \\ N h_i(\mathbf{x}) \mathbf{e}_i \end{pmatrix}, \tag{63}$$

where $\mathbf{e}_i$ is the $i$-th standard basis vector. Of course, this form of the oracle will not necessarily be a good choice for specific applications.

In particular, as discussed in Section 1.2 and in the corollaries of our main theorems, our results will apply in their full generality and they will improve deterministic complexity as long as $L \le \sqrt{N} L_F$, where $L$ is the Lipschitz constant corresponding to stochastic oracle in view of Assumption 1 and $L_F$ is for the full operator. However, it is not clear that the generic choice in (63) will satisfy this requirement. Therefore, one should be careful to design suitable oracles depending on the particular structure of the problem to ensure complexity improvements. We refer to Section 1.1 for a detailed comparison with related works.

## Appendix E. Numerical experiments

In this section, we provide preliminary empirical evidence[2] on how variance reduced methods for VIs perform in practice. By no means, this report is exhaustive, but only an illustration for showing *(i)* variance reduction helps in practice compared to deterministic methods and *(ii)* our approach is not only more general in theory but also offers practical advantages compared to the previous approach in (Carmon et al., 2019).

We focus on matrix games with simplex constraints in the Euclidean and entropic setups. In the Euclidean step, we use the projection to simplex from (Condat, 2016). We compare deterministic extragradient (EG), existing variance-reduced method (Carmon et al., 2019) (EG-Car+19) and proposed Alg. 1 and Alg. 2. To distinguish from the Euclidean case, we write 'MP' instead of 'EG'

2. Code can be found in https://github.com/ymalitsky/vr_for_vi

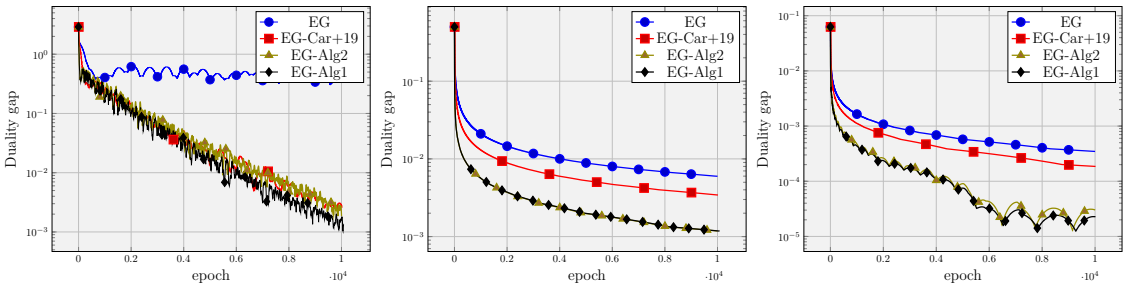


Figure 1: Euclidean setup. left: policeman and burglar matrix (Nemirovski, 2013), middle, right: two test matrices given in (Nemirovski et al., 2009, Section 4.5).

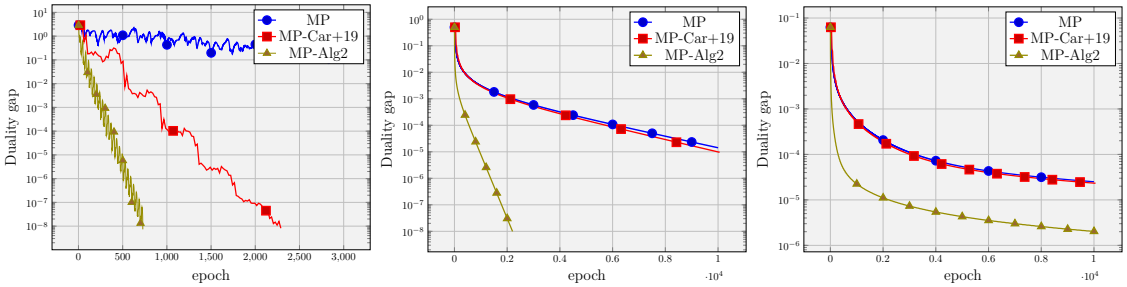


Figure 2: Entropic setup. The same matrices in Figure 1 used in the same arrangement.

for all algorithms. We have chosen three test problems used in the literature (Nemirovski, 2013; Nemirovski et al., 2009).

For all problems, we fix $m = n = 500$ and use the largest step sizes allowed by theory. In particular, EG uses $1/L_F$, where $L_F$ is the Lipschitz constant of the overall operator $F$. We also use the reported parameters from (Carmon et al., 2019) for EG-Car+19. In the Euclidean case, by tracing the proof of (Carmon et al., 2019, Proposition 2), we observed that one can improve the step size from $\eta = \frac{\alpha}{10L^2}$ to $\eta = \frac{\alpha}{4L^2}$, where $\alpha$ is defined to be $\frac{L\sqrt{m+n}}{\sqrt{\text{nnz}(A)}}$ therein. Therefore, we use the improved step size for EG-Car+19 for experiments with Euclidean setup. However, in the Bregman setup, we did not find a way to improve the step size of EG-Car+19, so we use the reported one.

In our methods, we use the parameters from Remarks 10 and 12. For performance measure, we use duality gap, which can be simply computed as $\max_i(A\mathbf{x})_i - \min_j(A^\top \mathbf{y})_j$ due to simplex constraints. Cost of computing one $F$ is counted as an epoch, and the cost of stochastic oracles are counted accordingly to match the overall cost.

We report the results in Figures 1 and 2. We see that variance reduced variants consistently outperform deterministic EG in all cases, as predicted in theory. Within variance reduced methods, due to the small step sizes of EG-Car+19, except the first dataset in the Euclidean setup, we observe our algorithms to also outperform EG-Car+19. Especially in the Bregman setting, the difference is noticeable since the analysis of EG-Car+19 requires smaller step sizes.