arXiv:2102.08352v1 [math.OC] 16 Feb 2021

Stochastic Variance Reduction for Variational Inequality Methods

Ahmet Alacaoglu* Yura Malitsky†

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 exactly 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.

1 Introduction

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.

Canonical examples of VIs are min-max problems

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}),$$

where the corresponding F and g are defined to be

$$F(\mathbf{z}) = \begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{pmatrix}, \qquad g(\mathbf{z}) = \delta_{\mathcal{X}}(\mathbf{x}) + \delta_{\mathcal{Y}}(\mathbf{y}),$$

with $\delta_{\mathcal{X}}$ denoting the indicator function of the set $\mathcal{X}$. In this case, (1) constitutes a necessary first order optimality condition, which for a convex-concave f is also sufficient. Informally, the operator F takes care of the smooth part, while g covers the non-smooth one.

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 [Nes05; Nem04; CP11; EZC10]. The second has been started by machine learning community, where solving nonconvex-nonconcave saddle point problems became of paramount importance [Gid+19; GM18; Mer+19]. Additionally, VIs have applications in game theory, control theory, and differential equations (see [FP07]).

A common structure encountered in min-max problems is that the operator F can be written as a finite-sum: $F = F_1 + \cdots + F_N$. Variance reduction techniques use this specific form to improve theoretical complexity of deterministic methods in minimization. Existing results on variance reduction for saddle point problems show that this 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.

*EPFL, Switzerland, email: ahmet.alacaoglu@epfl.ch
†LiU, Sweden, email: yurii.malitskyi@liu.se

	Assumptions	Complexity
[Kor76; Tse00; Nem04; MT20] EG/MP, FBF, FoRB	F is monotone	$\mathcal{O}\left(\mathsf{Cost} \times \frac{NL_F}{\varepsilon}\right)$
[Car+19] EG/MP	F is monotone + $\mathbf{z} \mapsto \langle F(\mathbf{z}) + \tilde{\nabla} g(\mathbf{z}), \mathbf{z} - \mathbf{u}\rangle$ is cvx. $\forall \mathbf{u}$; or bounded domains	$\mathcal{O}\left(\mathsf{Cost} \times \left(N + \frac{\sqrt{N}L}{\varepsilon}\right)\right)$
[AMC20] FoRB	F is monotone	$\mathcal{O}\left(\mathsf{Cost} \times \left(N + \frac{NL}{\varepsilon}\right)\right)$
This paper EG/MP, FBF, FoRB	F is monotone	$\mathcal{O}\left(\mathsf{Cost} \times \left(N + \frac{\sqrt{N}L}{\varepsilon}\right)\right)$

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.

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.

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 seems 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.

1.1 Related works

Variational inequalities. The standard choices for solving VIs have been methods such as extragradient (EG)/Mirror-Prox (MP) [Kor76; Nem04], forward-backward-forward (FBF) [Tse00], dual extrapolation [Nes07] or reflected gradient/forward-reflected-backward (FoRB) [Mal15; MT20][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, sublinear convergence rate $\mathcal{O}(\varepsilon^{-1})$ for monotone problems and linear rate of convergence for strongly monotone problems.

Variance reduction. Variance reduction approach 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" $\mathbb{E}\left\|\mathbf{g_k} - \nabla f(\mathbf{x}_k)\right\|^2$ compared to $\mathbb{E}\left\|\nabla f_i(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\right\|^2$ 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 [JZ13; DBL14]; for a more thorough list of reference see a recent review [Gow+20].

[1] In the unconstrained setting, this method is also known as Optimistic Mirror Descent (OMD) or Optimistic Gradient Descent Ascent (OGDA) [RS13; Das+18]

Variance reduction and VIs. One does not need to be meticulous to quickly encounter 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 we have already mentioned, saddle point reformulations often come to rescue.

The work [BB16] 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 bilinear problems, the complexity in [BB16] improves the deterministic method in the strongly monotone setting. [Cha+19] 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 [BB16].

Strong monotonicity may seem like a fine assumption, similar to strong convexity in minimization. While algorithmically it is indeed true, in applications with min-max, the former is far less frequent. Therefore, it is crucial to remove this assumption.

An influential work in this direction is by [Car+19], 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 $\mathbf{u}$ [Car+19, Corollary 1] or that domain is bounded [Car+19, 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.

Finally, there are other works that did not improve complexity but introduced new ideas. An algorithm similar in spirit to ours is due to [AMC20], where variance reduction is applied to FoRB. This algorithm was the first to converge under only monotonicity, but it did not improve complexity upon deterministic methods. Several works studied VI methods in the stochastic approximation setting and showed slower rates with decreasing step sizes [Mis+20; Böh+20] or increasing mini-batch sizes [Ius+17; Boţ+19; CS19].

1.2 Outline of results and comparisons

Complexity and ε-accurate solution. We say that a point $\bar{\mathbf{z}}$ is an ε-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 operations needed to reach an ε-accurate solution.

General case. We consider the problem (1) with $F = \sum_{i=1}^{N} F_i$ where F is monotone and L_F-Lipschitz. For comparison, let us assume that per iteration cost of stochastic methods is Cost and for deterministic methods, it is $\mathsf{Cost} \times N$. In this setting, our variance reduced variants of EG, FBF, and FoRB (Corollary 2.6, Corollary 4.3, Corollary 4.8) have complexity $\mathcal{O}\left(\mathsf{Cost} \times \left(N + \sqrt{N} L \varepsilon^{-1}\right)\right)$ compared to the deterministic methods with complexity $\mathcal{O}\left(\mathsf{Cost} \times \left(N L_F \varepsilon^{-1}\right)\right)$.

Our methods improve upon deterministic variants as long as $L \leq \sqrt{N} L_F$ (see Section 2.1 for the definitions of Lipschitz constants). This is a similar improvement over deterministic complexity, as accelerated variance reduction does for minimization problems [WS16; All17].

To our knowledge, the only precedent with a result similar to ours is the work [Car+19], where additional assumptions were required (see Section 1.1), complexity had an additional logarithmic term and a more complicated algorithm was needed.

Bilinear problems. When we focus on bilinear problems (Section 5.1), 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 4.3, Corollary 4.8).

	Rate & Complexity	Convergence of iterates
Euclidean setup		
EG: Section 2	Theorem 2.5, Corollary 2.6	Theorem 2.3
FBF, FoRB: Section 4	Corollary 4.3, Corollary 4.8	Theorem 4.2, Theorem 4.5
Bregman setup		
MP: Section 3	Theorem 3.6, Corollary 3.7	—

Table 2: Structure of the paper

In both settings this complexity was first obtained in [Car+19]. Our results generalize the set of problems where this complexity applies due to less assumptions and also use more practical/simpler algorithms (see Section 6 for an empirical comparison). We also remark that our variance reduced Mirror-Prox (see Algorithm 2) is different from the Mirror-Prox variant in [Car+19, Algorithm 1, Algorithm 2].

How to read the paper? We summarize the main results in Table 2. We recommend a reader, who wants a quick grasp of the idea, to refer to Section 2. This should be sufficient for understanding our main technique. The extension to Bregman case is technical in nature and noticing the reason for using a double loop algorithm in this case requires a good deal of understanding of proposed analysis.

For the most general case with Bregman distances, a reader can skip Section 2 without losing much and go directly to Section 3. We kept Section 2 for a clearer exposition of the main ideas via a simpler algorithm and analysis. We tried to make the sections self-contained and the proofs isolated: convergence rate and convergence of iterates are separated.

Finally, one can read Section 4 right after Section 2 to see how the same ideas give rise to variance reduced FBF and FoRB algorithms with similar guarantees and ability to solve monotone inclusions. In this section, we also illustrate how to obtain linear rate of convergence with strong convexity. Section 5 clarifies how to apply our developments to specific problems such as matrix games and linearly constrained optimization. Most of the proofs are given with the corresponding results; remaining proofs are deferred to Section 8.

Practical guide. We give the parameters recommended in practice in Remark 2.1 for Algorithm 1, Remark 3.1 for Algorithm 2, Remark 4.1 for Algorithm 3. 3 and Remark 4.4 for Algorithm 4. These parameters are optimized to obtain the best complexity in terms of dependence to problem dimensions (and not dependence to constants) and we use them in our numerical experiments in Section 6. For convenience we also specify the updates in the important case of matrix games with entropic setup in Section 5.1.2.

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 $\operatorname{dom} g = \{\mathbf{z} : g(\mathbf{z}) < +\infty\}$ and the following prox-inequality is standard

$$\bar{\mathbf{z}} = \operatorname{prox}_g(\mathbf{z}) \quad \iff \quad \langle \bar{\mathbf{z}} - \mathbf{z}, \mathbf{x} - \bar{\mathbf{z}} \rangle \geq 0, \qquad \forall \mathbf{x} \in \mathcal{Z}. \tag{3}$$

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

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_ξ that is unbiased $F(\mathbf{z}) = \mathbb{E}\left[F_\xi(\mathbf{z})\right]$ and L-Lipschitz in mean:

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

Finite sum. Suppose F has a finite sum representation $F = \sum_{i\in[N]} F_i$, 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 $L_F \le \sum_{i\in[N]} L_i$. On one hand, $\sum_{i\in[N]} L_i$ 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. Let us consider some suitable stochastic oracles.

The first one is the most straightforward. Let ξ be a random variable distributed uniformly on $[N]$. Define the stochastic oracle $F_\xi = NF_i$. Clearly, it is unbiased and

$$\mathbb{E}\,\|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|^2 = \sum_{i=1}^N N\|F_i(\mathbf{u}) - F_i(\mathbf{v})\|^2 \le N\sum_{i=1}^N L_i^2\|\mathbf{u}-\mathbf{v}\|^2.$$

Hence, $L = \sqrt{N\sum_{i\in[N]} L_i^2}$.

If, however, we sample ξ according to "importance" of F_i, then we can improve the above estimate. Namely, let ξ be a random variable such that $\Pr\{\xi = i\} = p_i = \frac{L_i}{\sum_{j\in[N]} L_j}$ for $i \in [N]$ and define the stochastic oracle $F_\xi = \frac{1}{p_i}F_i$. It is easy to verify that this oracle is L-Lipschitz with $L = \sum_{i\in[N]} L_i$. Indeed,

$$\mathbb{E}\,\|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|^2 = \sum_{i=1}^N \frac{1}{p_i}\|F_i(\mathbf{u}) - F_i(\mathbf{v})\|^2 \le \sum_{i=1}^N \frac{1}{p_i}L_i^2\|\mathbf{u}-\mathbf{v}\|^2 = \Big(\sum_{i=1}^N L_i\Big)^2\|\mathbf{u}-\mathbf{v}\|^2.$$

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 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.

2.2 Extragradient with variance reduction

The classical stochastic variance reduced gradient (SVRG) [JZ13] 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 [KHR20; Hof+15] 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 the same guarantees, but the latter variant is slightly simpler to analyze and implement.

We present the loopless version of extragradient with variance reduction, in Algorithm 1. Every iteration requires two stochastic oracles F_ξ and one F with probability p. Hence, we set p in the sequel as a small number. Parameter α is the key in establishing 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 α. Therefore, the specific form of $\bar{\mathbf{z}}_k$ is important. Later, we will see that with $\alpha = 1-p$, Algorithm 1 has $\mathcal{O}\left(\mathsf{Cost} \times \frac{\sqrt{N}L}{\varepsilon}\right)$ complexity.

Remark 2.1. For running algorithm in practice, we suggest $p = \frac{2}{N}$, $\alpha = 1 - p$, and $\tau = \frac{0.99\sqrt{p}}{L}$. However, specific problem may require a more careful examination of "optimal" parameters (see Section 5.1).

Algorithm 1 Extragradient with variance reduction (loopless)

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

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 ξ_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

It is interesting to note that by eliminating all randomness, Algorithm 1 reduces to the classic extragradient method. In fact, if $p = 1$ and $F_\xi = F$, then $\mathbf{w}_k = \mathbf{z}_k$, hence, $\bar{\mathbf{z}}_k = \mathbf{z}_k$, and the updates become

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

2.3 Analysis

For the iterates $(\mathbf{z}_k)$, $(\mathbf{w}_k)$ of Algorithm 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 Φ_k naturally arises in our analysis as the Lyapunov function.

Lemma 2.2. *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 Algorithm 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 $\sum_{k=0}^{\infty}\left((1-\alpha)\mathbb{E}\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2 + \mathbb{E}\|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2\right) \le \frac{1}{1-\gamma}\Phi_0(\mathbf{z}_*)$.

Proof. By prox-inequality (3) applied to the definitions of $\mathbf{z}_{k+1}$ and $\mathbf{z}_{k+1/2}$, we have that for all $\mathbf{z}$,

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

We sum two inequalities and arrange to get

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

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

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

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 $\mathbb{E}_k\left[F(\mathbf{w}_k) + F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k)\right] = F(\mathbf{z}_{k+1/2})$ to obtain

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

and

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

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

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

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

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

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

We proceed to show the almost sure convergence of the sequence by using Lemma 2.2 and standard arguments from the literature of stochastic optimization methods [Ber11; CP15]. To show a global convergence, we need F_ξ to be Lipschitz for all ξ. For a finite sum example it follows automatically from Assumption 1.

Theorem 2.3. *Let Assumption 1 hold, F_ξ be Lipschitz for all ξ, $\alpha \in [0,1)$, $p \in (0,1]$, 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 Algorithm 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 $\operatorname{dom} g$. As proven in [Nes07, Lemma 1], this restricted version of the gap is a valid measure as long as $\mathcal{C}$ contains any solution $\mathbf{z}_* \in \mathsf{Sol}$. Since we work in probabilistic setting, naturally our convergence measure will be based on $\mathbb{E}[\mathrm{Gap}]$.

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 [Nem+09].

Lemma 2.4. *Let $\mathcal{F} = (\mathcal{F}_k)_{k\geq 0}$ 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 $\mathcal{C} \subset \mathcal{Z}$,*

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

Theorem 2.5. *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 denote the cost of computing one $F_\xi(\cdot)$ as Cost, and the cost of computing $F(\cdot)$ as $\mathsf{Cost}\times N$. For a finite sum example, as in Section 2.1, this is the most natural assumption.

Corollary 2.6. *Let the conjecture of Theorem 2.5 hold. Then the average total complexity (see Remark 2.7) of Algorithm 1 to reach ε-accuracy is bounded by $\mathcal{O}\left(\mathsf{Cost}\times(pN+2)\left(1+\frac{L}{\sqrt{p}\varepsilon}\right)\right)$. In particular, for $p = \frac{2}{N}$ it is $\mathcal{O}\left(\mathsf{Cost}\times\frac{\sqrt{N}L}{\varepsilon}\right)$.*

Remark 2.7. One might notice that the result in Table 1 has an additional $\mathsf{Cost}\times N$. The difference is mainly due to the notion of complexity in cases of "loopless" (single loop) vs "looped" (double loop) algorithms. For Algorithm 1, since per iteration cost is $\mathsf{Cost}\times(pN+1)$ *in expectation*, the result is "average" total complexity: *expected number of iterations to get a small expected gap*. On the other hand, Algorithm 2 has a fixed cost per iteration, thus, it gives a more standard notion of complexity: *number of iterations to get a small expected gap*.

Remark 2.8. To see the justification for the choice of $\alpha = 1-p$, consider the proof with any choice of α. The resulting bound will be $\mathcal{O}\left(\frac{1}{\sqrt{1-\alpha}} + \frac{\sqrt{1-\alpha}}{p}\right)$. Then $\alpha = 1-p$ optimizes it in terms of p dependence.

Proof of Theorem 2.5. As we have already mentioned, when all randomness is eliminated, that is $F_\xi = F$ and $p = 1$, Algorithm 1 reduces to the extragradient. In that case, the convergence rate $\mathcal{O}(1/K)$ would follow almost immediately from the proof of Lemma 2.2. In a stochastic setting the proof is more subtle and we have to rely on Lemma 2.4 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 2.2 before getting (10). In particular, using (6) and (7) in (5) gives

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

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

Now, we set $\alpha = 1-p$. We want to rewrite (12) 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 (12) 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

$$\begin{aligned}2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \le \max_{\mathbf{z}\in\mathcal{C}}\Phi_0(\mathbf{z})+\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\left(e_1(\mathbf{z},k)+e_2(\mathbf{z},k)\right)\right]\\ -\mathbb{E}\sum_{k=0}^{K-1}\left(\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2+p\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2\right)\\ +2\tau\mathbb{E}\sum_{k=0}^{K-1}\left[\langle F_{\xi_k}(\mathbf{w}_k)-F_{\xi_k}(\mathbf{z}_{k+1/2}),\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\rangle\right]\end{aligned}\tag{13}$$

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 $\mathbf{z}_{k+1/2}\mapsto\langle F(\mathbf{z}),\mathbf{z}_{k+1/2}-\mathbf{z}\rangle$, and convexity of g.

The tower property, the estimation from (9), and $1-\alpha=p$ applied on (13) imply

$$2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \le \max_{\mathbf{z}\in\mathcal{C}}\Phi_0(\mathbf{z})+\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\left(e_1(\mathbf{z},k)+e_2(\mathbf{z},k)\right)\right].\tag{14}$$

Therefore, the proof will be complete upon deriving an upper bound for the second term on RHS. We will instantiate Lemma 2.4 twice for bounding this term. First, for $e_1(\mathbf{z},k)$ we set in Lemma 2.4,

$$\mathcal{F}_k=\sigma(\xi_0,\dots,\xi_{k-1},\mathbf{w}_k),\quad \tilde{\mathbf{x}}_0=\mathbf{z}_0,\quad \mathbf{u}_{k+1}=2\tau\left(\left[F_{\xi_k}(\mathbf{z}_{k+1/2})-F_{\xi_k}(\mathbf{w}_k)\right]-\left[F(\mathbf{z}_{k+1/2})-F(\mathbf{w}_k)\right]\right),$$

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

$$\begin{aligned}\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}e_1(\mathbf{z},k)\right]&=\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\langle\mathbf{u}_{k+1},\mathbf{z}\rangle\right]-\mathbb{E}\left[\sum_{k=0}^{K-1}\langle\mathbf{u}_{k+1},\mathbf{z}_{k+1/2}\rangle\right]=\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\langle\mathbf{u}_{k+1},\mathbf{z}\rangle\right]\\ &\le\max_{\mathbf{z}\in\mathcal{C}}\frac{1}{2}\|\mathbf{z}_0-\mathbf{z}\|^2+\frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{u}_{k+1}\|^2\\ &\le\max_{z\in\mathcal{C}}\frac{1}{2}\|\mathbf{z}_0-\mathbf{z}\|^2+2\tau^2L^2\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2,\end{aligned}\tag{15}$$

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

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

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 2.4

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

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

$$\begin{aligned}\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1} e_2(\mathbf{z},k)\right] &= 2\,\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}}\sum_{k=0}^{K-1}\langle \mathbf{u}_{k+1}, z\rangle\right]\\ &\le \max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2 + \sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{u}_{k+1}\|^2\\ &= \max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2 + p(1-p)\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2, \end{aligned} \tag{16}$$

where the inequality follows from Lemma 2.4 and the second equality from the derivation

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

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

Combining (15), (16), and (14), we finally arrive at

$$\begin{aligned}2\tau K\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^K)\right] \le \max_{\mathbf{z}\in\mathcal{C}}\Phi_0(\mathbf{z}) &+ \max_{\mathbf{z}\in\mathcal{C}}\frac{1}{2}\|\mathbf{z}_0-\mathbf{z}\|^2 + 2\tau^2L^2\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2\\ &+ \max_{\mathbf{z}\in\mathcal{C}}\|\mathbf{z}_0-\mathbf{z}\|^2 + p(1-p)\sum_{k=0}^{K-1}\mathbb{E}\,\|\mathbf{z}_{k+1}-\mathbf{w}_k\|^2\end{aligned} \tag{17}$$

We have to estimate terms under the sum:

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

where the first inequality in (18) uses Lemma 2.2 and $1-\alpha = p$.

Now we will use that $\mathbf{w}_0 = \mathbf{z}_0$ and, hence, $\Phi_0(\mathbf{z}) = (2-p)\|\mathbf{z}_0-\mathbf{z}\|^2 \le 2\|\mathbf{z}_0-\mathbf{z}\|^2$ in (17). 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

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

In particular, 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 2.6. In average each iteration costs $pN\mathsf{Cost} + 2\mathsf{Cost} = (pN+2)\mathsf{Cost}$. To reach ε-accuracy we need $\left\lceil\mathcal{O}\left(\frac{L}{\sqrt{p}\varepsilon}\right)\right\rceil$ iterations. Hence, the total average complexity is $\mathcal{O}\left(\frac{\mathsf{Cost}\times(pN+2)L}{\sqrt{p}\varepsilon}\right)$. Finally, the optimal choice $p = \frac{2}{N}$ results in $\mathcal{O}\left(\frac{\mathsf{Cost}\times\sqrt{N}L}{\varepsilon}\right)$ complexity. ■

3 Bregman setup

3.1 Preliminaries

In this section, we will 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

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

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

$$\langle \nabla h(\mathbf{x}) - \nabla h(\mathbf{y}), \mathbf{z} - \mathbf{x} \rangle = D(\mathbf{z}, \mathbf{y}) - D(\mathbf{z}, \mathbf{x}) - D(\mathbf{x}, \mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{Z}. \tag{19}$$

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_ξ of F with as small L as possible. Moreover, the proof of Lemma 2.2 indicates that in k-th iteration we need Lipschitzness only for already known two iterates. Hence, following [GK95; Car+19], in contrast to Algorithm 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.

Definition 1. We say that F has a stochastic oracle F_ξ 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) $\mathbb{E}_{\xi \sim Q_{\mathbf{u},\mathbf{v}}}\left[\|F_\xi(\mathbf{u}) - F_\xi(\mathbf{v})\|_*^2\right] \leq L^2 \|\mathbf{u} - \mathbf{v}\|^2$.

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 [Car+19, 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.

For brevity we introduce the new set of assumptions. It is important to remark that Assumption 2 is not a restriction of Assumption 1: every item is either the same or more general.

Assumption 2.

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

(ii) The function $g \in \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is proper convex lsc.

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

(iv) The operator F has a stochastic oracle F_ξ that is variable L-Lipschitz in mean (see Definition 1).

3.2 Mirror-Prox with variance reduction

In this setting, we can simply adjust the steps of Algorithm 1 and correspondingly the analysis of Lemma 2.2. However, to show a convergence rate, double randomization in Algorithm 1 causes technical complications. For this reason, in the Bregman setup we propose a double loop variant of Algorithm 1, similar to the classical SVRG [JZ13]. Our algorithm can be seen as a variant of Mirror-Prox [Nem04] with variance reduction. Now it should be clear that Algorithm 1 is a randomized version of Algorithm 2 with $p = \frac{1}{K}$ and a particular choice $D(\mathbf{z}, \mathbf{z}') = \frac{1}{2}\|\mathbf{z} - \mathbf{z}'\|_2^2$.

Algorithm 2 Mirror-prox with variance reduction

Input: Step size τ, $\alpha \in (0,1)$, $K > 0$. 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**
    $\mathbf{z}_{k+1/2}^s = \operatorname{argmin}_{\mathbf{z}} \left\{ g(\mathbf{z}) + \langle F(\mathbf{w}^s), \mathbf{z} \rangle + \frac{\alpha}{\tau} D(\mathbf{z}, \mathbf{z}_k^s) + \frac{1-\alpha}{\tau} D(\mathbf{z}, \bar{\mathbf{w}}^s) \right\}$.
    Fix distribution $Q_{\mathbf{z}_{k+1/2}^s, \mathbf{w}^s}$ and sample ξ_k^s according to it
    $\mathbf{z}_{k+1}^s = \operatorname{argmin}_{\mathbf{z}} \left\{ g(\mathbf{z}) + \langle F(\mathbf{w}^s) + F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s), \mathbf{z} \rangle + \frac{\alpha}{\tau} D(\mathbf{z}, \mathbf{z}_k^s) + \frac{1-\alpha}{\tau} D(\mathbf{z}, \bar{\mathbf{w}}^s) \right\}$.
  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

Compared to Algorithm 1, $\mathbf{w}^s$ serves the same purpose as $\mathbf{w}_k$: the snapshot point in the language of SVRG [JZ13]. Since we have two loops in this case, we get $\mathbf{w}^s$ by averaging, again, similar to SVRG for non-strongly convex optimization [Red+16; AY16]. 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.

Remark 3.1. For running algorithm in practice, we suggest $K = \frac{N}{2}$, $\alpha = 1 - \frac{1}{K}$, and $\tau = \frac{0.99\sqrt{p}}{L}$.

3.3 Analysis

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

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

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

Lemma 3.2. *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 $\mathbf{z}$,

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

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

$$\Theta_{k+1/2}^s(\mathbf{z}) = \langle F(\mathbf{z}_{k+1/2}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z} \rangle + g(\mathbf{z}_{k+1/2}^s) - g(\mathbf{z}), \tag{20}$$

$$e(\mathbf{z}, s, k) = \tau \langle F(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F(\mathbf{w}^s) + F_{\xi_k^s}(\mathbf{w}^s), \mathbf{z}_{k+1/2}^s - \mathbf{z} \rangle. \tag{21}$$

$$\delta(s,k) = \tau \langle F_{\xi_k^s}(\mathbf{w}^s) - F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s), \mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s \rangle - \frac{1}{2} \|\mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\|^2 - \frac{1-\alpha}{2} \|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2 \tag{22}$$

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

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

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

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

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

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

Remark 3.4. We will use Lemma 3.3(i) and Lemma 3.3(iii) for proving the convergence rate. On the other hand, Lemma 3.3(ii) can be used to derive subsequential convergence, which we do not include for brevity.

Proof of Lemma 3.3. Applying Lemma 3.2 to $\mathbf{z}_{k+1/2}^s$ update, with $\mathbf{z} = \mathbf{z}_{k+1}^s$, we have

$$\tau\left(g(\mathbf{z}_{k+1}^s) - g(\mathbf{z}_{k+1/2}^s) + \langle F(\mathbf{w}^s), \mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\rangle\right) \geq D(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) \\ + \alpha\Big(D(\mathbf{z}_{k+1/2}^s, \mathbf{z}_k^s) - D(\mathbf{z}_{k+1}^s, \mathbf{z}_k^s)\Big) + (1-\alpha)\Big(D(\mathbf{z}_{k+1/2}^s, \bar{\mathbf{w}}^s) - D(\mathbf{z}_{k+1}^s, \bar{\mathbf{w}}^s)\Big). \quad (23)$$

Applying Lemma 3.2 to $\mathbf{z}_{k+1}^s$ update with a general $\mathbf{z} \in \mathcal{Z}$, we have

$$\tau\left(g(\mathbf{z}) - g(\mathbf{z}_{k+1}^s) + \langle F(\mathbf{w}^s) + F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s), \mathbf{z} - \mathbf{z}_{k+1}^s\rangle\right) \geq D(\mathbf{z}, \mathbf{z}_{k+1}^s) \\ + \alpha\Big(D(\mathbf{z}_{k+1}^s, \mathbf{z}_k^s) - D(\mathbf{z}, \mathbf{z}_k^s)\Big) + (1-\alpha)\Big(D(\mathbf{z}_{k+1}^s, \bar{\mathbf{w}}^s) - D(\mathbf{z}, \bar{\mathbf{w}}^s)\Big). \quad (24)$$

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

$$D(\mathbf{u}, \bar{\mathbf{w}}^s) - D(\mathbf{v}, \bar{\mathbf{w}}^s) = \frac{1}{K}\sum_{j=1}^{K}\left(D(\mathbf{u}, \mathbf{z}_j^{s-1}) - D(\mathbf{v}, \mathbf{z}_j^{s-1})\right). \quad (25)$$

Summing up (23) and (24) and using (25), we obtain

$$\tau\Big(g(\mathbf{z}) - g(\mathbf{z}_{k+1/2}^s) + \langle F(\mathbf{w}^s) + F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s), \mathbf{z} - \mathbf{z}_{k+1/2}^s\rangle\Big) \geq D(\mathbf{z}, \mathbf{z}_{k+1}^s) - \alpha D(\mathbf{z}, \mathbf{z}_k^s) \\ + \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}_{k+1/2}^s, \mathbf{z}_j^{s-1}) - \frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}, \mathbf{z}_j^{s-1}) + D(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) \\ + \tau\langle F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s), \mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\rangle. \quad (26)$$

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

$$\frac{1-\alpha}{K}\sum_{j=1}^{K} D(\mathbf{z}_{k+1/2}^s, \mathbf{z}_j^{s-1}) \geq \frac{1-\alpha}{K}\sum_{j=1}^{K}\frac{1}{2}\|\mathbf{z}_{k+1/2}^s - \mathbf{z}_j^{s-1}\|^2 \geq \frac{1-\alpha}{2}\|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2, \quad (27)$$

$$D(\mathbf{z}_{k+1}^s, \mathbf{z}_{k+1/2}^s) \geq \frac{1}{2}\|\mathbf{z}_{k+1}^s - \mathbf{z}_{k+1/2}^s\|^2. \quad (28)$$

By using (20), (27), and (28) in (26), we deduce

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

where we have defined the last term as $e(\mathbf{z},s,k)$ (see (21)). We sum this inequality over k to obtain the result in (i).

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

$$\begin{aligned}\tau\mathbb{E}_{s,k}\langle F_{\xi^s_k}(\mathbf{w}^s) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2}), \mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\rangle &\le \mathbb{E}_{s,k}\left[\frac{\tau^2}{2}\|F_{\xi^s_k}(\mathbf{w}^s) - F_{\xi^s_k}(\mathbf{z}^s_{k+1/2})\|^2_* + \frac{1}{2}\|\mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\|^2\right]\\ &\le \frac{(1-\alpha)\gamma^2}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2 + \frac{1}{2}\mathbb{E}_{s,k}\|\mathbf{z}^s_{k+1} - \mathbf{z}^s_{k+1/2}\|^2, \quad (29)\end{aligned}$$

since $\tau^2L^2 = (1-\alpha)\gamma^2$. We take expectation of (26), plug in $\mathbf{z} = \mathbf{z}_*$; use (8), (29), (27), and (28) to get

$$\mathbb{E}_{s,k}\left[D(\mathbf{z}_*,\mathbf{z}^s_{k+1})\right] \le \alpha D(\mathbf{z}_*,\mathbf{z}^s_k) + \frac{1-\alpha}{K}\sum_{j=1}^K D(\mathbf{z}_*,\mathbf{z}^{s-1}_j) + \frac{(1-\alpha)(\gamma^2-1)}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2. \quad (30)$$

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

$$\mathbb{E}_{s,0}D(\mathbf{z}_*,\mathbf{z}^s_{k+1}) \le \mathbb{E}_{s,0}\Big[\alpha D(\mathbf{z}_*,\mathbf{z}^s_k) + \frac{1-\alpha}{K}\sum_{j=1}^K D(\mathbf{z}_*,\mathbf{z}^{s-1}_j) - \frac{(1-\alpha)(1-\gamma^2)}{2}\|\mathbf{z}^s_{k+1/2} - \mathbf{w}^s\|^2\Big]. \quad (31)$$

Summing this inequality over $k = 0,\ldots,K-1$ 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 2.4.

Lemma 3.5. *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 $\mathcal{C} \subset \operatorname{dom} g$*

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

Theorem 3.6. *Let Assumption 1 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*

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

Corollary 3.7. *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 Algorithm 2 to reach ε-accuracy is $\mathcal{O}\left(\mathsf{Cost} \times \left(N + \frac{L\sqrt{N}}{\varepsilon}\right)\right)$. In particular, if $\tau = \frac{\sqrt{1-\alpha}}{3L} = \frac{\sqrt{2}}{3\sqrt{N}L}$, the total complexity is $\mathsf{Cost} \times \left(2N + \frac{43\sqrt{N}L}{\varepsilon}\max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0)\right)$.*

Proof of Theorem 3.6. We will start with the result of Lemma 3.3 and proceed similar to Theorem 2.5. Since $\mathbf{z}_0^{s+1} = \mathbf{z}_K^s$, we use definition of $\Phi^s(\mathbf{z})$, and sum the inequality in Lemma 3.3(i) over s to obtain

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

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

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

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 (29) can be applied here with $\delta(s,k)$ defined as (22)). Therefore,

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

We next bound the second term on RHS, similar to the proof of Theorem 2.5. For $s \in \{0,\dots,S-1\}$ and $k \in \{0,\dots,K-1\}$, set $\mathcal{F}_k^s = \sigma(\mathbf{z}_{1/2}^0,\dots,\mathbf{z}_{K-1/2}^0,\dots,\mathbf{z}_{1/2}^s,\dots,\mathbf{z}_{k+1/2}^s)$, $\mathbf{u}_{k+1}^s = \tau[F(\mathbf{w}^s) - F_{\xi_k^s}(\mathbf{w}^s) - F(\mathbf{z}_{k+1/2}^s) + F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s)]$, which help us write

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

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

$$\begin{aligned}
\mathbb{E}\left[\max_{\mathbf{z}\in\mathcal{C}} \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} e(\mathbf{z},k)\right] &\le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \tau^2\mathbb{E}\|F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s) + F(\mathbf{w}^s) - F(\mathbf{z}_{k+1/2}^s)\|_*^2 \\
&\le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} 4\tau^2\mathbb{E}\|F_{\xi_k^s}(\mathbf{z}_{k+1/2}^s) - F_{\xi_k^s}(\mathbf{w}^s)\|_*^2 \qquad (32) \\
&\le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \sum_{s=0}^{S-1}\sum_{k=0}^{K-1} 4\tau^2L^2\mathbb{E}\|\mathbf{z}_{k+1/2}^s - \mathbf{w}^s\|^2 \qquad (33) \\
&\le \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z},\mathbf{z}_0) + \frac{8\tau^2L^2}{(1-\alpha)(1-\gamma^2)}\Phi^0(\mathbf{z}_*), \qquad (34)
\end{aligned}$$

where (32) 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, (33) is by variable Lipschitzness of F_ξ, and the last step is by Lemma 3.3. 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 $\tau^2L^2 = (1-\alpha)\gamma^2$ we have

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

Proof of Corollary 3.7. As $\alpha = 1 - \frac{1}{K}$, it holds that $\alpha + K(1-\alpha) = 1 - \frac{1}{K} + 1 \leq 2$. With this, from Theorem 3.6 it follows

$$\begin{aligned}\mathbb{E}\left[\mathrm{Gap}(\mathbf{z}^S)\right] &\leq \frac{1}{\tau K S}\left(1 + \left(1 + \frac{8\gamma^2}{1-\gamma^2}\right)(\alpha + K(1-\alpha)\right) \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0) \\ &\leq \frac{L}{\sqrt{K}\gamma S}\left(3 + \frac{16\gamma^2}{1-\gamma^2}\right) \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0) = \mathcal{O}\left(\frac{L}{\sqrt{N}S}\right). \end{aligned} \tag{35}$$

One epoch requires one evaluation of F and $2K$ of F_ξ, therefore in total we have $(N + 2K)\mathsf{Cost} = 2N\mathsf{Cost}$. To reach ε accuracy, we need $\left\lceil \mathcal{O}\left(\frac{L}{\sqrt{N}\varepsilon}\right)\right\rceil$ epochs. Hence, the final complexity is $\mathcal{O}\left(\mathsf{Cost} \times \left(N + \frac{L\sqrt{N}}{\varepsilon}\right)\right)$. Now, by setting $\gamma = \frac{1}{3}$ in (35), 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 $\mathsf{Cost} \times \left(2N + \frac{43\sqrt{N}L}{\varepsilon} \max_{\mathbf{z}\in\mathcal{C}} D(\mathbf{z}, \mathbf{z}_0)\right)$. ■

Remark 3.8. Because we work with general norms, we had to use in (34) 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 3.7 leads to $\mathsf{Cost} \times \left(2N + \frac{13\sqrt{N}L}{\varepsilon} \max_{\mathbf{z}\in\mathcal{C}} \|\mathbf{z} - \mathbf{z}_0\|^2\right)$ total complexity for the Euclidean setting.

4 Extensions

In this section, we show how to obtain the variance reduced versions of two other operator splitting methods: forward-backward-forward (FBF) [Tse00] and forward-reflected-backward (FoRB) [MT20] 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

$$\mathbf{z}_* \in \mathcal{Z} \text{ such that } 0 \in (F + G)(\mathbf{z}^*), \tag{36}$$

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 (36) 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 a stochastic oracle F_ξ that is unbiased $F(\mathbf{z}) = \mathbb{E}_\xi\left[F_\xi(\mathbf{z})\right]$ and L-Lipschitz in mean:

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

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: $\mathbb{E}[\cdot|\sigma(\xi_0, \ldots, \xi_{k-1}, \mathbf{w}_k)] = \mathbb{E}_k[\cdot]$ and $\mathbb{E}[\cdot|\sigma(\xi_0, \ldots, \xi_k, \mathbf{w}_k] = \mathbb{E}_{k+1/2}[\cdot]$. 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 (36) becomes a VI (1) and $J_G = \mathrm{prox}_g$.

Algorithm 3 FBF with variance reduction

Input: Probability $p \in (0,1]$, probability distribution Q, step size τ, $\alpha \in (0,1)$. 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 ξ_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

4.1 Forward-Backward-Forward with variance reduction

Forward-backward-forward (FBF) algorithm was introduced by Tseng in [Tse00]. 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 [Böh+20].

Remark 4.1. For running algorithm in practice, we suggest $p = \frac{2}{N}$, $\alpha = 1-p$, and $\tau = \frac{0.99\sqrt{p}}{L}$.

We have essentially the same convergence result as for extragradient in Section 2.3. We keep the same notation as therein and recall the definition of Φ_k for convenience

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

Theorem 4.2. *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 Algorithm 3 and any $\mathbf{z}_* \in \mathsf{Sol}$, it holds that*

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

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

Proof of Theorem 4.2. 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

$$\langle \mathbf{z}_{k+1/2} - \bar{\mathbf{z}}_k + \tau\left(F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k) + F(\mathbf{w}_k)\right), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle - \langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle \geq 0. \tag{37}$$

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

$$\begin{aligned} 2\langle \mathbf{z}_{k+1} - \bar{\mathbf{z}}_k, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle &= 2\langle \mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle + 2\langle \mathbf{z}_{k+1/2} - \bar{\mathbf{z}}_k, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle \\ &= \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 + \|\mathbf{z} - \mathbf{z}_{k+1/2}\|^2 - \|\mathbf{z} - \mathbf{z}_{k+1}\|^2 + 2\langle \mathbf{z}_{k+1/2} - \bar{\mathbf{z}}_k, \mathbf{z} - \mathbf{z}_{k+1/2}\rangle \\ &= \|\mathbf{z}_{k+1} - \mathbf{z}_{k+1/2}\|^2 - \|\mathbf{z} - \mathbf{z}_{k+1}\|^2 + \alpha\|\mathbf{z} - \mathbf{z}_k\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}\|^2 \\ &\quad - \alpha\|\mathbf{z}_{k+1/2} - \mathbf{z}_k\|^2 - (1-\alpha)\|\mathbf{z}_{k+1/2} - \mathbf{w}_k\|^2. \end{aligned} \tag{38}$$

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

$$2\tau\mathbb{E}_k\left[\langle F_{\xi_k}(\mathbf{z}_{k+1/2}) - F_{\xi_k}(\mathbf{w}_k) + F(\mathbf{w}_k), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle\right] = 2\tau\mathbb{E}_k\left[\langle F(\mathbf{z}_{k+1/2}), \mathbf{z} - \mathbf{z}_{k+1/2}\rangle\right]. \tag{39}$$

We use (38) and (39) in (37) to obtain

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

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

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

This derives the first result, which is the analogue of Lemma 2.2. To show almost sure convergence, we basically follow the proof of Theorem 2.3. First, using Robbins-Siegmund theorem and [CP15, Proposition 2.3] as in Theorem 2.3, we obtain that there exists a probability 1 set Ξ 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 (53), we use the definitions of $\mathbf{z}_{k+1/2}$, resolvent, and $\mathbf{z}_{k+1}$ to obtain

$$\begin{aligned}\mathbf{z}_{k+1}(\theta) - \bar{\mathbf{z}}_k(\theta) + \tau\left(F_{\xi_k}(\mathbf{w}_k(\theta)) - F_{\xi_k}(\mathbf{z}_{k+1/2}(\theta))\right) + \tau\left(F(\mathbf{z}_{k+1/2}(\theta)) - F(\mathbf{w}_k(\theta))\right) \\ \in \tau\left(F+G\right)(\mathbf{z}_{k+1/2}(\theta)),\end{aligned} \tag{40}$$

to show that $\tilde{\mathbf{z}}(\theta) \in (F+G)^{-1}(0)$. In particular, we use that F_ξ is Lipschitz for all ξ, $\mathbf{z}_{k+1} - \bar{\mathbf{z}}_k \to 0$, and $\mathbf{z}_{k+1/2} - \mathbf{w}_k \to 0$ almost surely. We use the same arguments as the proof of Theorem 2.3 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 4.3. *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 ε-accurate solution is $\mathcal{O}\left(\mathsf{Cost} \times \frac{\sqrt{N}L}{\epsilon}\right)$.*

4.2 Forward-reflected-backward with variance reduction: revisited

In a similar spirit to FBF, but using a different idea, [MT20] 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 [RS13; Das+18] and in some particular cases is equivalent to Popov's method [Pop80]. Later, in [AMC20], the authors suggested the most straightforward variance reduction modification of FoRB by combining FoRB and loopless SVRG [KHR20]. 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 [AMC20], 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 [AMC20; Zha21].

Remark 4.4. For running algorithm in practice, we suggest $p = \frac{2}{N}$, $\alpha = 1 - p$, and $\tau = \frac{0.99\sqrt{p(1-p)}}{L}$.

Algorithm 4 FoRB with variance reduction

Input: Probability $p \in (0,1]$, probability distribution Q, step size τ, $\alpha \in (0,1)$. 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 ξ_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

Lyapunov function here is slightly more complicated than previous sections. In particular, it is given as

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

Theorem 4.5. *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 Algorithm 4 and any $\mathbf{z}_* \in \mathsf{Sol}$, it holds that $\Phi_k(\mathbf{z}_*)$ is nonnegative and*

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

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

Remark 4.6. Note that again when randomness is null, $F_\xi = F$ and $p = 1$, Algorithm 4 reduces to the original FoRB algorithm. Moreover, with $\alpha = \frac{1}{2}$ we will have the same guarantee as in [MT20].

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

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

$$\langle \mathbf{z}_{k+1} - \bar{\mathbf{z}}_k + \tau\left[F(\mathbf{w}_k) - F_{\xi_k}(\mathbf{w}_{k-1}) + F_{\xi_k}(\mathbf{z}_k)\right], \mathbf{z} - \mathbf{z}_{k+1}\rangle - \tau\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}_{k+1}\rangle \geq 0. \tag{41}$$

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

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

Second, as we derived in (6),

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

Substituting the last two estimates into (41), we obtain

$$\begin{aligned}
&\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_{k+1}) - F(\mathbf{w}_k), \mathbf{z} - \mathbf{z}_{k+1}\rangle + 2\tau\langle F(\mathbf{z}) - F(\mathbf{z}_{k+1}), \mathbf{z} - \mathbf{z}_{k+1}\rangle \\
&\leq \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}\|^2 + 2\tau\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z} - \mathbf{z}_k\rangle \\
&\quad + 2\tau\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}_k - \mathbf{z}_{k+1}\rangle - \alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 - (1-\alpha)\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2.
\end{aligned} \tag{42}$$

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

$$\begin{aligned}
&\mathbb{E}_k \left[\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_{k+1}) - F(\mathbf{w}_k), \mathbf{z} - \mathbf{z}_{k+1}\rangle + (1-\alpha)\|\mathbf{z}_{k+1} - \mathbf{w}_k\|^2\right] \\
&\leq \alpha\|\mathbf{z}_k - \mathbf{z}\|^2 + (1-\alpha)\|\mathbf{w}_k - \mathbf{z}\|^2 + 2\tau\langle F(\mathbf{z}_k) - F(\mathbf{w}_{k-1}), \mathbf{z} - \mathbf{z}_k\rangle \\
&\quad + 2\tau\, \mathbb{E}_k \left[\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}_k - \mathbf{z}_{k+1}\rangle - \alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2\right]. \qquad (43)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_k[2\tau\langle F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1}), \mathbf{z}_k - \mathbf{z}_{k+1}\rangle - \alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2] \\
&\leq \mathbb{E}_k \left[\frac{\tau^2}{\alpha\gamma}\|F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1})\|^2 + \alpha\gamma\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 - \alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2\right] \\
&\leq \frac{(1-\alpha)\gamma}{L^2}\, \mathbb{E}_k \|F_{\xi_k}(\mathbf{z}_k) - F_{\xi_k}(\mathbf{w}_{k-1})\|^2 - (1-\gamma)\alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\
&\leq (1-\alpha)\gamma\|\mathbf{z}_k - \mathbf{w}_{k-1}\|^2 - (1-\gamma)\alpha\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2. \qquad (44)
\end{aligned}$$

Adding (11) and (44) to (43), 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 4.2. The only difference is that instead of (40), we have

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

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

Remark 4.7. Even though we will set the parameters α, p, τ by optimizing complexity, we observe that the requirements in Theorem 4.5 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 [AMC20].

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 [AMC20] needs to make use of multiple Young's inequalities. In particular, we use $\mathbf{z}_*$ as an anchor point in (11), whereas [AMC20] 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 4.3, we give the complexity of the algorithm for solving VI in the spirit of Section 2.3.1.

Corollary 4.8. *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 ε-accurate solution is $\mathcal{O}\left(\mathsf{Cost} \times \frac{\sqrt{N}L}{\epsilon}\right)$.*

4.3 Linear convergence

In this section, we illustrate how to obtain linear convergence of Algorithm 1 for solving VI (1) when g is μ-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 4.9. *Let Assumption 1 hold, g be μ-strongly convex, and $\mathbf{z}_*$ be the solution of (1). If we set $\alpha = 1 - p$ and $\tau = \frac{\sqrt{p}}{2L}$ in Algorithm 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 will 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 2.2 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

$$(1-p+\tau\mu)\,\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2\right]+\mathbb{E}_k\left[\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2\right] \leq (1-p)\|\mathbf{z}_k-\mathbf{z}_*\|^2+\|\mathbf{w}_k-\mathbf{z}_*\|^2 \\ -\frac{1}{2}\left(p\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right]\right). \quad (45)$$

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 (45) and that $c\leq\frac{\sqrt{p}\mu}{2L}=\tau\mu$ gives us

$$\left(1-p+\frac{c}{3}\right)\mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_*\|^2\right]+\left(1+\frac{c}{3}\right)\mathbb{E}_k\left[\|\mathbf{w}_{k+1}-\mathbf{z}_*\|^2\right] \leq (1-p)\|\mathbf{z}_k-\mathbf{z}_*\|^2+\|\mathbf{w}_k-\mathbf{z}_*\|^2 \\ -\frac{1}{2}\left(p\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right)+\frac{4c}{3}\left(\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2+\mathbb{E}_k\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right). \quad (46)$$

By our choice of c, we have $\frac{4c}{3}\leq\frac{p}{2}$ and, therefore, the second line of (46) 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

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

Corollary 4.10. *Let $p=\frac{2}{N}$, $\tau=\frac{\sqrt{p}}{2L}$. The total average complexity in this case is $\mathcal{O}\left(\mathsf{Cost}\times\left(N+\frac{\sqrt{N}L}{\mu}\right)\log\frac{1}{\epsilon}\right)$.*

Proof. The ε-accuracy will be reached after $\mathcal{O}(\log\frac{1}{\varepsilon}/\log(1+\frac{c}{3}))$ iterations. This will yield 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 $\log\left(\varepsilon^{-1}\right)$. $\blacksquare$

Remark 4.11. In this case, Algorithm 1 has complexity $\mathcal{O}\left(\mathsf{Cost}\times\left(N+\frac{\sqrt{N}L}{\mu}\right)\log\frac{1}{\varepsilon}\right)$, compared to the deterministic methods $\mathcal{O}\left(\mathsf{Cost}\times\frac{NL_F}{\mu}\log\frac{1}{\varepsilon}\right)$. This complexity recovers the previously obtained result in [BB16], where our advantage is having algorithmic parameters independent of μ. Compared to [Car+19, Proposition 6] where the complexity is $\mathcal{O}\left(\mathsf{Cost}\times\left(N+\frac{\sqrt{N}L}{\mu}\log\frac{1}{\varepsilon}\right)\right)$, our result is slightly worse. On the other hand, our assumptions are more general (see Section 1.1).

5 Applications

5.1 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 we denote a number of its non-zero entries by $\text{nnz}(A)$; 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 $\|A\| = \sigma_{\max}(A)$, $\|A\|_{\text{Frob}} = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\sum_{i=1}^{\text{rank}(A)} \sigma_i(A)^2}$, 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 $A_{:j}$.

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

$$F(\mathbf{z}) = F(\mathbf{x},\mathbf{y}) = \begin{pmatrix} A^\top\mathbf{y} \\ -A\mathbf{x}\end{pmatrix}, \qquad g(\mathbf{z}) = g_1(\mathbf{x}) + g_2(\mathbf{y}). \tag{47}$$

5.1.1 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 (47) when $g_1(\mathbf{x}) = f(\mathbf{x})$ and $g_2(\mathbf{y}) = \langle b, \mathbf{y}\rangle$.

We will instantiate Algorithm 1 for this problem. To make our presentation clearer, we consider only the most common scenario when $\text{nnz}(A) > m + n$. In this setting, deterministic methods (extragradient, FBF, FoRB, etc.) solve (49) with $\mathcal{O}\left(\text{nnz}(A)\|A\|\varepsilon^{-1}\right)$ total complexity. As we see in the sequel, variance reduced methods provide us $\mathcal{O}\left(\text{nnz}(A) + \sqrt{\text{nnz}(A)(m+n)}\|A\|_{\text{Frob}}\varepsilon^{-1}\right)$ total complexity. We now describe the definition of F_ξ 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_ξ can be computed as

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

Oracle 2. The second stochastic oracle is slightly more complicated, since it is iteration-dependent as [Car+19]. We use 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 Algorithm 2 we define a distribution $Q(\mathbf{z}_{k+1/2}^s, \mathbf{w}^s)$ and sample ξ according to it.

Clearly, as before, F_ξ 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 5.1.2.

Complexity. We suppose that computing proximal operators prox_{g_1}, prox_{g_2} can be done efficiently in $\tilde{\mathcal{O}}(m+n)$ complexity. Our result in Theorem 2.5 stated that Algorithm 1 has the rate $\mathcal{O}\left(\frac{L}{\sqrt{p}K}\right)$. Given that the expected cost of each iteration is $\mathcal{O}\left(p\,\text{nnz}(A) + m + n\right)$, setting $p = \frac{m+n}{\text{nnz}(A)}$ gives us the average total complexity

$$\tilde{\mathcal{O}}\left(m + n + \frac{\sqrt{\text{nnz}(A)(m+n)}\|A\|_{\text{Frob}}}{\varepsilon}\right). \tag{48}$$

It is easy to see that Algorithm 2 has the complexity $\tilde{\mathcal{O}}\left(\text{nnz}(A) + \frac{\sqrt{\text{nnz}(A)(m+n)}\|A\|_{\text{Frob}}}{\varepsilon}\right)$ if we set $K = \left\lceil \frac{\text{nnz}(A)}{m+n} \right\rceil$. Please see Remark 2.7 for the reason of slight difference in complexities.

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 $\|A\|_{\text{Frob}} \leq \sqrt{\text{rank}(A)}\|A\|$, the bound in (48) improves that of deterministic method. Since dual problem is separable, one can use randomized primal-dual methods to exploit this property [Cha+18; AFC19]. However, the worst-case complexity in this case stays the same as the deterministic method, which can be worked out from the proof of Theorem 4.9 in [AFC19].

Finally, we remark that the analysis in [Car+19, 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 compared to us.

5.1.2 Matrix games

The problem in this case is written as

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \langle A\mathbf{x}, \mathbf{y}\rangle, \tag{49}$$

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 (47), 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 [Car+19]. 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 5.1.1 and we obtain the same complexity as (48). The same discussions as Section 5.1.1 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 (49) is known as a zero sum game. In this case, deterministic algorithms formulated with a specific Bregman distance (given below) have $\mathcal{O}\left(\text{nnz}(A)\|A\|_{\max}\varepsilon^{-1}\right)$ 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 Algorithm 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 (47) 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 5.1.1 for the Euclidean case, but with adjustment to the ℓ_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_ξ is variable $\|A\|_{\max}$-Lipschitz in view of Definition 1. Indeed, we have

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

Similarly, in every iteration of Algorithm 2 we define a distribution $Q(\mathbf{z}_{k+1/2}^s, \mathbf{w}^s)$ and sample ξ_k^s according to it. This stochastic oracle was already used in [GK95] and used extensively after that, see [NN13; CHW12] and references therein. In [Car+19] this oracle was called "sampling from the difference".

Complexity. In this case, the complexity of deterministic algorithms (Mirror Prox, FoRB) is $\mathcal{O}\left(\operatorname{nnz}(A)\|A\|_{\max}\varepsilon^{-1}\right)$. Our result in Corollary 3.7 stated that Algorithm 2 has the rate $\mathcal{O}\left(\frac{L}{\sqrt{KS}}\right)$. Given that the cost of each epoch of Algorithm 2 is $\mathcal{O}\left(\operatorname{nnz}(A) + K(m+n)\right)$, setting $K = \left\lceil \frac{\operatorname{nnz}(A)}{m+n} \right\rceil$ gives us the total complexity

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

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

Updates. For concreteness we specify updates in lines 4–6 of Algorithm 2. Let $\mathbf{w}^s = (\mathbf{u}, \mathbf{v})$, $\bar{\mathbf{w}}^s = (\bar{\mathbf{u}}^s, \bar{\mathbf{v}}^s)$.

$$\nabla h_1(\mathbf{x}^s_{k+1/2}) = \alpha \nabla h_1(\mathbf{x}^s_k) + (1-\alpha)\nabla h_1(\bar{\mathbf{u}}^s) - \tau A^\top \mathbf{v}^s$$
$$\nabla h_2(\mathbf{y}^s_{k+1/2}) = \alpha \nabla h_2(\mathbf{y}^s_k) + (1-\alpha)\nabla h_2(\bar{\mathbf{v}}^s) + \tau A \mathbf{u}^s$$

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

$$\Pr\{\xi = (i,j)\} = r_i c_j, \quad r_i = \frac{|y^s_{k+1/2,i} - v^s_i|}{\|\mathbf{y}^s_{k+1/2} - \mathbf{v}^s\|_1}, \quad c_j = \frac{|x^s_{k+1/2,i} - u^s_i|}{\|\mathbf{x}^s_{k+1/2} - \mathbf{u}^s\|_1}$$

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

$$\begin{aligned}\nabla h_1(\mathbf{x}^s_{k+1}) &= \alpha \nabla h_1(\mathbf{x}^s_k) + (1-\alpha)\nabla h_1(\bar{\mathbf{u}}^s) - \tau A^\top \mathbf{v}^s - \frac{\tau}{r_i} A_{i:}(y^s_{k+1/2,i} - v^s_i) \\ &= \nabla h_1(\mathbf{x}^s_{k+1/2}) - \tau A_{i:} \|\mathbf{y}^s_{k+1/2} - \mathbf{v}^s\| \operatorname{sign}(y^s_{k+1/2,i} - v^s_i) \\ \nabla h_2(\mathbf{y}^s_{k+1}) &= \nabla h_2(\mathbf{y}^s_{k+1/2}) + \tau A_{:j} \|\mathbf{x}^s_{k+1/2} - \mathbf{u}^s\| \operatorname{sign}(x^s_{k+1/2,j} - u^s_j)\end{aligned}$$

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

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

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

5.2 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

$$F_i(\mathbf{z}) = \begin{pmatrix} \nabla f(\mathbf{x}) + N y_i \nabla h_i(\mathbf{x}) \\ N h_i(\mathbf{x}) \mathbf{e}_i \end{pmatrix}, \tag{50}$$

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 (50) 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.

6 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

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

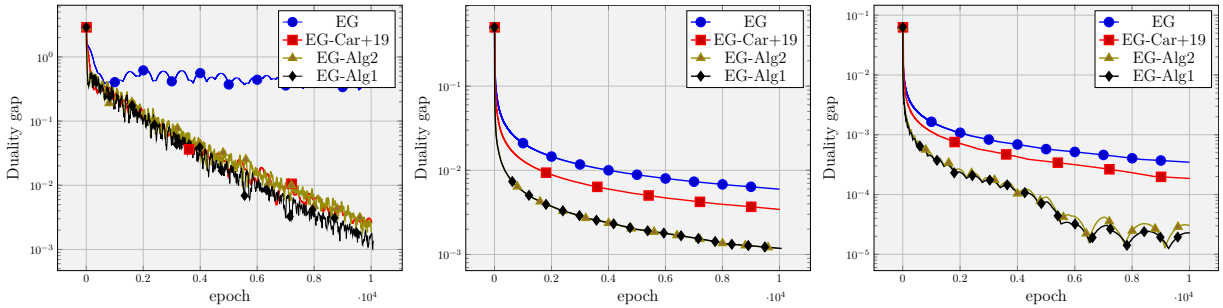

Figure 1: Euclidean setup. left: policeman and burglar matrix [Nem13], middle, right: two test matrices given in [Nem+09, Section 4.5].

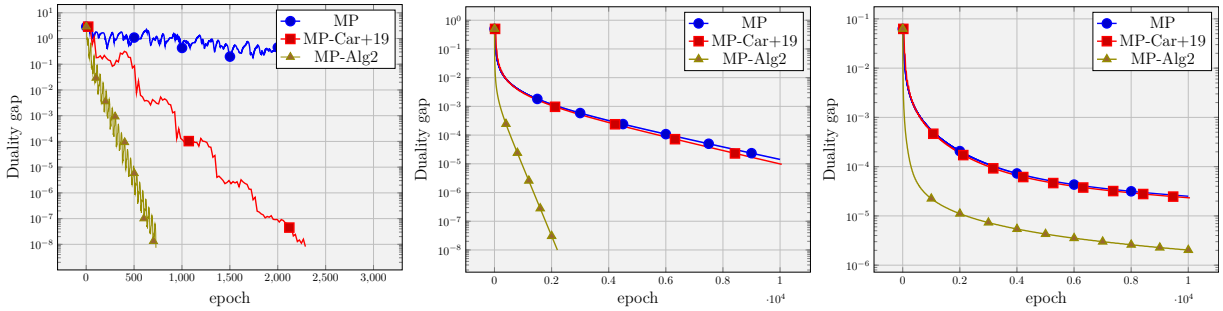

Figure 2: Entropic setup. The same matrices in Figure 1 used in the same arrangement.

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 [Car+19].

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 [Con16]. We compare deterministic extragradient (EG), existing variance-reduced method [Car+19] (EG-Car+19) and proposed Algorithm 1 and Algorithm 2. To distinguish from the Euclidean case, we write 'MP' instead of 'EG' for all algorithms. We have chosen three test problems used in the literature [Nem13; Nem+09].

For all problems, we fix dimension $m = n = 500$ 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 [Car+19] for EG-Car+19. In the Euclidean case, by tracing the proof of [Car+19, Proposition 2], we observed that one can improve the step size from $\eta = \frac{\alpha}{10L^2}$ to $\eta = \frac{\alpha}{4L^2}$, where α is defined to be $\frac{L\sqrt{m+n}}{\sqrt{\mathrm{nnz}(A)}}$ 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 2.1 and 3.1. 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.

7 Conclusions

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

Sparsity. An important consideration in practice is to adapt to sparsity of the data. The recent work by [Car+20] built on the algorithm in [Car+19] and improved the complexity for matrix games in Euclidean setup, for sparse data, by using specialized data structures. We believe that these techniques can also be used in our algorithms.

Lower bounds. Our results match or improve the best-known complexities given in [BB16; Car+19]. However, there are no matching lower bounds to show optimality of these results. For example, in minimization, optimality of accelerated variance reduction is known due to tight lower bounds for finite sums in [WS16]. For min-max problems, a recent result in [OX19] established tight lower bounds for deterministic methods. Obtaining similar results in the finite sum min-max setting is an important open problem.

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.

8 Appendix

Proof of Theorem 2.3. By the proof of Lemma 2.2, without removing the term $-\alpha\|\mathbf{z}_{k+1/2}-\mathbf{z}_k\|^2$ in (7), we have

$$\mathbb{E}_k\left[\Phi_{k+1}(\mathbf{z}_*)\right] \le \Phi_k(\mathbf{z}_*) - (1-\gamma)\left((1-\alpha)\|\mathbf{z}_{k+1/2}-\mathbf{w}_k\|^2 + \mathbb{E}_k\left[\|\mathbf{z}_{k+1}-\mathbf{z}_{k+1/2}\|^2\right]\right) - \alpha\|\mathbf{z}_{k+1/2}-\mathbf{z}_k\|^2. \quad (51)$$

By Robbins-Siegmund theorem [RS71, 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 $Q = \operatorname{diag}(\alpha,\dots,\alpha,\frac{1-\alpha}{p},\dots,\frac{1-\alpha}{p})$. Then applying [CP15, 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 Ξ, 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 Ξ with $\mathbb{P}(\Xi)=1$, such that

$$\forall\theta\in\Xi \text{ and } \forall \mathbf{z}_*\in\mathsf{Sol},\quad \alpha\|\mathbf{z}_k(\theta)-\mathbf{z}_*\|^2 + \frac{1-\alpha}{p}\|\mathbf{w}_k(\theta)-\mathbf{z}_*\|^2 \text{ converges.} \quad (52)$$

Moreover, by taking total expectation on (51), 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 Ξ' 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 $\mathbf{z}_{k+1}$,

$$\langle \mathbf{z}_{k+1}(\theta)-\bar{\mathbf{z}}_k(\theta)+\tau F(\mathbf{w}_k(\theta))-\tau F_{\xi_k}(\mathbf{z}_{k+1/2}(\theta))+\tau F_{\xi_k}(\mathbf{w}_k(\theta)), \mathbf{z}-\mathbf{z}_{k+1}(\theta)\rangle + \tau g(\mathbf{z})-\tau g(\mathbf{z}_{k+1}(\theta))\ge 0,\quad \forall \mathbf{z}\in\mathcal{Z}. \quad (53)$$

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_ξ is Lipschitz for all ξ 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 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 (52) 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 Sol. ■

Proof of Lemma 3.2. 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

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

■

Proof of Lemma 2.4. 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 Lemma 3.5. Define for each $s \geq 0$ and for $k \in \{0, \ldots, K-1\}$,

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

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 manipulate 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

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

We sum this inequality over $k = 0, \ldots, K-1$ and $s = 0, \ldots, S-1$, take maximum, use $\mathbf{x}_0^{s+1} = \mathbf{x}_K^s$ and the same derivations as at the end of the proof of Lemma 2.4 to show $\sum_{s=0}^{S-1}\sum_{k=0}^{K-1} \mathbb{E}\left[\langle \mathbf{u}_{k+1}^s, \mathbf{x}_k^s\rangle\right] = 0$. ■

Acknowledgments

The work of A. Alacaoglu has received funding from 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

[AFC19] A. Alacaoglu, O. Fercoq, and V. Cevher. *On the convergence of stochastic primal-dual hybrid gradient*. 2019. arXiv: 1911.00799.

[AMC20] A. Alacaoglu, Y. Malitsky, and V. Cevher. "Forward-reflected-backward method with variance reduction". In: *under review* (2020).

[All17] Z. Allen-Zhu. "Katyusha: The first direct acceleration of stochastic gradient methods". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 8194–8244.

[AY16] Z. Allen-Zhu and Y. Yuan. "Improved SVRG for non-strongly-convex or sum-of-non-convex objectives". In: *International conference on machine learning*. PMLR. 2016, pp. 1080–1089.

[BB16] P. Balamurugan and F. Bach. "Stochastic variance reduction methods for saddle-point problems". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1416–1424.

[Ber11] D. P. Bertsekas. "Incremental proximal methods for large scale convex optimization". In: *Mathematical Programming* 129.2 (2011), p. 163.

[Böh+20] 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*. 2020. arXiv: 2006.09033.

[Boţ+19] R. I. Boţ, P. Mertikopoulos, M. Staudigl, and P. T. Vuong. *Forward-backward-forward methods with variance reduction for stochastic variational inequalities*. 2019. arXiv: 1902.03355.

[Car+19] Y. Carmon, Y. Jin, A. Sidford, and K. Tian. "Variance reduction for matrix games". In: *Advances in Neural Information Processing Systems*. 2019, pp. 11377–11388.

[Car+20] Y. Carmon, Y. Jin, A. Sidford, and K. Tian. *Coordinate Methods for Matrix Games*. 2020. arXiv: 2009.08447.

[Cha+18] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schonlieb. "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications". In: *SIAM Journal on Optimization* 28.4 (2018), pp. 2783–2808.

[CP11] A. Chambolle and T. Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145.

[Cha+19] 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*. 2019, pp. 391–401.

[CHW12] K. L. Clarkson, E. Hazan, and D. P. Woodruff. "Sublinear optimization for machine learning". In: *Journal of the ACM (JACM)* 59.5 (2012), pp. 1–49.

[CP15] P. L. Combettes and J.-C. Pesquet. "Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping". In: *SIAM Journal on Optimization* 25.2 (2015), pp. 1221–1248.

[Con16] L. Condat. "Fast projection onto the simplex and the ℓ_1 ball". In: *Mathematical Programming* 158.1 (2016), pp. 575–585.

[CS19] 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*. 2019. arXiv: 1904.11076.

[Das+18] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. "Training GANs with Optimism". In: *International Conference on Learning Representations*. 2018.

[DBL14] 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*. 2014, pp. 1646–1654.

[EZC10] 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". In: *SIAM Journal on Imaging Sciences* 3.4 (2010), pp. 1015–1046.

[FP07] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[GM18] I. Gemp and S. Mahadevan. *Global convergence to the equilibrium of GANs using variational inequalities*. 2018. arXiv: 1808.01531.

[Gid+19] 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.

[Gow+20] R. M. Gower, M. Schmidt, F. Bach, and P. Richtarik. "Variance-reduced methods for machine learning". In: *Proceedings of the IEEE* 108.11 (2020), pp. 1968–1983.

[GK95] M. D. Grigoriadis and L. G. Khachiyan. "A sublinear-time randomized approximation algorithm for matrix games". In: *Operations Research Letters* 18.2 (1995), pp. 53–58.

[Hof+15] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. "Variance reduced stochastic gradient descent with neighbors". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2305–2313.

[Ius+17] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. "Extragradient method with variance reduction for stochastic variational inequalities". In: *SIAM Journal on Optimization* 27.2 (2017), pp. 686–724.

[JZ13] R. Johnson and T. Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in Neural Information Processing Systems*. 2013, pp. 315–323.

[Kor76] G. Korpelevich. "The extragradient method for finding saddle points and other problems". In: *Ekon. Mat. Metody* 12 (1976), pp. 747–756.

[KHR20] 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: *Proceedings of the 31st International Conference on Algorithmic Learning Theory*. 2020, pp. 451–467.

[Mal15] Y. Malitsky. "Projected reflected gradient methods for monotone variational inequalities". In: *SIAM Journal on Optimization* 25.1 (2015), pp. 502–520.

[MT20] Y. Malitsky and M. K. Tam. "A forward-backward splitting method for monotone inclusions without cocoercivity". In: *SIAM Journal on Optimization* 30.2 (2020), pp. 1451–1472.

[Mer+19] 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.

[Mis+20] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. "Revisiting stochastic extragradient". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 4573–4582.

[Nem04] 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". In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251.

[Nem13] A. Nemirovski. *Mini-Course on Convex Programming Algorithms*. Lecture notes. 2013.

[Nem+09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609.

[Nes05] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Mathematical programming* 103.1 (2005), pp. 127–152.

[NN13] Y. Nesterov and A. Nemirovski. "On first order algorithms for ℓ_1/nuclear norm minimization". In: *Acta Numer* 22 (2013), pp. 509–575.

[Nes07] Y. Nesterov. "Dual extrapolation and its applications to solving variational inequalities and related problems". In: *Mathematical Programming* 109.2-3 (2007), pp. 319–344.

[OX19] Y. Ouyang and Y. Xu. "Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems". In: *Mathematical Programming* (2019), pp. 1–35.

[Pop80] L. D. Popov. "A modification of the Arrow-Hurwicz method for search of saddle points". In: *Mathematical notes of the Academy of Sciences of the USSR* 28.5 (1980), pp. 845–848.

[RS13] A. Rakhlin and K. Sridharan. "Online learning with predictable sequences". In: *Conference on Learning Theory*. PMLR. 2013, pp. 993–1019.

[Red+16] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. "Stochastic variance reduction for nonconvex optimization". In: *International conference on machine learning*. PMLR. 2016, pp. 314–323.

[RS71] H. Robbins and D. Siegmund. "A convergence theorem for non negative almost supermartingales and some applications". In: *Optimizing methods in statistics*. Elsevier, 1971, pp. 233–257.

[Tse00] P. Tseng. "A modified forward-backward splitting method for maximal monotone mappings". In: *SIAM Journal on Control and Optimization* 38.2 (2000), pp. 431–446.

[WS16] B. Woodworth and N. Srebro. "Tight complexity bounds for optimizing composite objectives". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 3646–3654.

[Zha21] H. Zhang. *Extragradient and extrapolation methods with generalized bregman distances for saddle point problems*. 2021. arXiv: 2101.09916.