SOLVING STOCHASTIC WEAK MINTY VARIATIONAL INEQUALITIES WITHOUT INCREASING BATCH SIZE

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper introduces a family of stochastic extragradient-type algorithms for a class of nonconvex-nonconcave problems characterized by the weak Minty variational inequality (MVI). Unlike existing results on extragradient methods in the monotone setting, employing diminishing stepsizes is no longer possible in the weak MVI setting. This has led to approaches such as increasing batch sizes per iteration which can however be prohibitively expensive. In contrast, our proposed methods involves two stepsizes and only requires one additional oracle evaluation per iteration. We show that it is possible to keep one fixed stepsize while it is only the second stepsize that is taken to be diminishing, making it interesting even in the monotone setting. Almost sure convergence is established and we provide a unified analysis for this family of schemes which contains a nonlinear generalization of the celebrated primal dual hybrid gradient algorithm.

1 INTRODUCTION

Stochastic first-order methods have been at the core of the current success in deep learning applications. These methods are mostly well-understood for minimization problems at this point. This is even the case in the nonconvex setting where there exists matching upper and lower bounds on the complexity for finding an approximately stable point (Arjevani et al., 2019).

The picture becomes less clear when moving beyond minimization into nonconvex-nonconcave minimax problems—or more generally nonmonotone variational inequalities. Even in the deterministic case, finding a stationary point is in general intractable (Daskalakis et al., 2021; Hirsch & Vavasis, 1987; Papadimitriou, 1994). This is in stark contrast with minimization where only global optimality is NP-hard.

An interesting nonmonotone class for which we *do* have efficient algorithms is characterized by the so called *weak Minty variational inequality* (MVI) (Diakonikolas et al., 2021). This problem class captures nontrivial structures such as attracting limit cycles and is governed by a parameter ρ whose negativity increases the degree of nonmonotonicity. It turns out that the stepsize γ for the exploration step in extragradient-type schemes lower bounds the problem class through $\rho > -\gamma/2$ (Pethick et al., 2022). In other words, it seems that we need to take γ large to guarantee convergence for a large class.

This large stepsize requirement is at the core of why the community has struggled to provide a stochastic variants for weak MVIs. The only known results effectively increase the batch size *at every iteration* (Diakonikolas et al., 2021, Thm. 4.5)—a strategy that would be prohibitively expensive in most machine learning applications. Pethick et al. (2022) proposed (SEG+) which attempts to tackle the noise by only diminishing the second stepsize. This suffices in the special case of unconstrained quadratic games but can fail even in the monotone case as illustrated in Figure 1. This naturally raises the following research question:

Can stochastic weak Minty variational inequilities be solved without increasing the batch size?

We resolve this open problem in the affirmative with a modification of stochastic extragradient called *bias-corrected stochastic extragradient* (BCSEG+), which only requires one additional first order oracle call, while crucially maintaining the fixed stepsize. Specifically, we make the following contributions:

(i) We show that it is possible to converge for weak MVI *without* increasing the batch size, by introducing a bias-correction term. The scheme introduces *no additional hyperparameters* and recovers the maximal range $\rho \in (-\gamma/2, \infty)$ of explicit deterministic schemes. The result carries over to another class of problem treated in Appendix F, which we call *negative* weak MVIs.

(ii) We generalize the result to a whole family of schemes that can treat constrained and regularized settings. First and foremost the class includes a generalization of the forward-backward-forward (FBF) algorithm of Tseng (2000) to stochastic weak MVIs. The class also contains a stochastic nonlinear extension of the celebrated primal dual hybrid gradient (PDHG) algorithm (Chambolle & Pock, 2011). Both methods are obtained as instantiations of the same template scheme, thus providing a unified analysis and revealing an interesting requirement on the update under weak MVI when only stochastic feedback is available.

(iii) We prove almost sure convergence under the classical Robbins-Monro stepsize schedule of the second stepsize. This provides a guarantee on the last iterate, which is especially important in the nonmonotone case, where average guarantees cannot be converted into a single candidate solution. Almost sure convergence is challenging already in the monotone case where even stochastic extragradient may not converge (Hsieh et al., 2020, Fig. 1).

2 Related work

Weak MVI Diakonikolas et al. (2021) was the first to observe that an extragradient-like scheme called extragradient+ (EG+) converges globally for weak MVIs with $\rho \in (-1/8L_F, \infty)$. This results was later tighten to $\rho \in (-1/2L_F, \infty)$ and extended to constrained and regularized settings in (Pethick et al., 2022) through an adaptive algorithm. A single-call variant has been analysed in Böhm (2022). Under the stronger condition of cohypomonotonicity, an inexact proximal point method was studied in Combettes & Pennanen (2004) and in the exact case a tight characterization was carried out by Bauschke et al. (2021). It was shown that acceleration is achievable for an extragradient-type scheme even in this nonmonotone case (Lee & Kim, 2021). Despite this array of positive results the stochastic case is largely untreated for weak MVIs. The only known result (Diakonikolas et al., 2021, Theorem 4.5) requires the batch size to be increasing. Similarly, the accelerated method in Lee & Kim (2021, Thm. 6.1) requires the variance of the stochastic oracle to decrease as $O(1/k)$.

Stochastic & monotone When monotone-like conditions holds the story is different since it suffices to take diminishing stepsizes. For bounded domain a stochastic version of Mirror-Prox was analysed in Juditsky et al. (2011). The boundedness assumption was later dropped in Mishchenko et al. (2020) by allowing a multi-point stochastic oracle as in Assumption II*(iii)*. There exists a multitude of extensions for monotone problems: Single-call stochastic methods for quasi-monotone problems has been covered in detail by Hsieh et al. (2019), the quasi strongly monotone case was studied in isolation in Gorbunov et al. (2022), variance reduction was applied to Halpern-type iteration when restricted to the monotone case (Cai et al., 2022), and restarted averaging was considered specifically for bilinear games (Li et al., 2022), just to mention a few.

Variance reduction The assumptions we make about the stochastic oracle in Section 3 are similar to what is found in the variance reduction literature (see for instance Alacaoglu & Malitsky (2021, Assumption 1) or Arjevani et al. (2019)). However, our use of the assumption are different in a crucial way. Whereas the variance reduction literature uses the stepsize $\gamma \propto 1/L_{\hat{F}}$ (see e.g. Alacaoglu & Malitsky (2021, Theorem 2.5)), we aim at using the much larger $\gamma \propto 1/L_F$. For instance in the special case of a finite sum problem of size N, the mean square smoothness constant $L_{\hat{F}}$ defined through Assumption III can be $\sqrt{N}$ times larger than L_F (see Appendix H for details). This would lead to a prohibitively strict requirement on the degree of allowed nonmonotonicity through the relationship $\rho > -\gamma/2$.

Bias-correction The idea of adding a correction term has also been exploited in minimization, specifically in the context of compositional optimization Chen et al. (2021). Due to their distinct problem setting it suffices to simply extend stochastic gradient descent (SGD), albeit under additional assumptions such as (Chen et al., 2021, Assumption 3). In our setting, however, SGD is not possible even when restricting ourselves to monotone problems.

3 Problem formulation and preliminaries

We are interesting in finding $z \in \mathbb{R}^n$ such that the following inclusion holds,

$$0 \in Tz := Az + Fz. \tag{3.1}$$

A wide range of machine learning applications can be cast as an inclusion. Most noticeable, a structured minimax problem can be reduced to (3.1) as shown in Section 8.1.

Assumption I. *In problem* (3.1)*,*

(i) The operator $F : \mathbb{R}^n \to \mathbb{R}^n$, to be defined, is L_F-Lipschitz with $L_F \in [0, \infty)$, i.e.,

$$\|Fz - Fz'\|_\Gamma \le L_F \|z - z'\|_{\Gamma^{-1}} \quad \forall z, z' \in \mathbb{R}^n. \tag{3.2}$$

where $\Gamma \in \mathbb{R}^{n\times n}$ is positive definite.

(ii) The operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a maximally monotone operator.

(iii) Weak Minty variational inequality (MVI) holds, i.e., there exists a nonempty set $\mathcal{S}^\star \subseteq \mathbf{zer}\, T$ such that for all $z^\star \in \mathcal{S}^\star$ and some $\rho \in (-\frac{1}{2L_F}, \infty)$

$$\langle v, z - z^\star \rangle_\Gamma \ge \rho \|v\|^2, \quad \text{for all } (z, v) \in \mathbf{gph}\, T. \tag{3.3}$$

Remark 1. In the unconstrained and smooth case ($A \equiv 0$), Assumption I*(iii)* reduces to $\langle Fz, z - z^\star \rangle \ge \rho \|Fz\|^2$ for all $z \in \mathbb{R}^n$. When $\rho = 0$ this condition reduces to star-monotonicity, while negative ρ makes the problem increasingly nonmonotone. Interestingly, the inequality is not symmetric and one may instead consider that the assumption holds for $-F$. Through this observation, Appendix F extends the reach of the extragradient-type algorithms developed for weak MVIs.

Stochastic oracle We assume that we cannot compute Fz easily, but instead we have access to the stochastic oracle $\hat{F}(z, \xi)$, which we assume is unbiased with bounded variance. We additionally assume that $z \mapsto \hat{F}(z, \xi)$ is $L_{\hat{F}}$ Lipschitz continuous in mean and that it can be simultaneously queried under the same randomness.

Assumption II. *For the operator $\hat{F}(\cdot, \xi) : \mathbb{R}^n \to \mathbb{R}^n$ the following holds.*

(i) Unbiased: $\mathbb{E}_\xi\left[\hat{F}(z, \xi)\right] = Fz$.

(ii) Bounded variance: $\mathbb{E}_\xi\left[\|\hat{F}(z, \xi) - \hat{F}(z, \xi)\|_\Gamma^2\right] \le \sigma_F^2$.

(iii) Two-point oracle: The stochastic oracle can be queried for any two points $z, z' \in \mathbb{R}^n$,

$$\hat{F}(z, \xi), \hat{F}(z', \xi) \quad \text{where} \quad \xi \sim \mathcal{P}. \tag{3.4}$$

Assumption III. *The operator $\hat{F}(\cdot, \xi) : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous in mean with $L_{\hat{F}} \in [0, \infty)$:*

$$\mathbb{E}_\xi\left[\|\hat{F}(z, \xi) - \hat{F}(z', \xi)\|_\Gamma^2\right] \leqslant L_{\hat{F}}^2 \|z - z'\|_{\Gamma^{-1}}^2 \quad \text{for all } z, z' \in \mathbb{R}^n. \tag{3.5}$$

Remark 2. The two latter assumptions are also common in the variance reduction literature (Fang et al., 2018; Nguyen et al., 2019; Alacaoglu & Malitsky, 2021), but in contrast with variance reduction we will not need knowledge of $L_{\hat{F}}$ to specify the algorithm, in which case the problem constant will only affect the complexity. Crucially, this decoupling of the stepsize from $L_{\hat{F}}$ will allow the proposed scheme to converge for a larger range of ρ in Assumption I*(iii)*. Finally, note that Assumption II*(iii)* commonly holds in machine learning applications, where usually the stochasticity is induced by the sampled mini-batch.

4 Method

To arrive at a stochastic scheme for weak MVI we first need to understand the crucial ingredients in the deterministic setting. For simplicity we will initially consider the unconstrained and smooth setting, i.e. $A \equiv 0$ in (3.1). The first component is taking the second stepsize α smaller as done in extragradient+ (EG+),

$$\begin{aligned} \bar{z}^k &= z^k - \gamma F z^k \\ z^{k+1} &= z^k - \alpha\gamma F \bar{z}^k \end{aligned} \tag{EG+}$$

Algorithm 1 (BCSEG+) Stochastic algorithm for problem (3.1) when $A \equiv 0$

REQUIRE $z^{-1} = \bar{z}^{-1} = z^0 \in \mathbb{R}^n$ $\alpha_k \in (0,1)$, $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$
REPEAT for $k = 0, 1, \ldots$ until convergence
1.1: Sample $\xi_k \sim \mathcal{P}$
1.2: $\bar{z}^k = z^k - \gamma \hat{F}(z^k, \xi_k) + (1 - \alpha_k)(\bar{z}^{k-1} - z^{k-1} + \gamma \hat{F}(z^{k-1}, \xi_k))$
1.3: Sample $\bar{\xi}_k \sim \mathcal{P}$
1.4: $z^{k+1} = z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k)$
RETURN z^{k+1}

where $\alpha \in (0,1)$. Convergence in weak MVI was first observed by Diakonikolas et al. (2021) and later tightened by Pethick et al. (2022), who characterized that smaller α allows for a larger range of the problem constant ρ. This is unproblematic for a stochastic scheme where usually the stepsize is taken diminishing regardless.

However, Pethick et al. (2022) also showed that the extrapolation stepsize γ plays a critical role for convergence under weak MVI. Specifically, they proved that a larger stepsize γ leads to a looser bound on the problem class through $\rho > -\gamma/2$ and that this bound is tight for (EG+). Unfortunately, picking γ large (e.g. as $\gamma = 1/L_F$) causes significant complications in the stochastic case where both stepsizes are usually taken to be diminishing as in the following scheme,

$$\begin{aligned} \bar{z}^k &= z^k - \beta_k \gamma \hat{F}(z^k, \xi_k) \quad \text{with} \quad \xi_k \sim \mathcal{P} \\ z^{k+1} &= z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k) \quad \text{with} \quad \bar{\xi}_k \sim \mathcal{P} \end{aligned} \tag{SEG}$$

where $\alpha_k = \beta_k = 1/k$. Even with a two-timescale variant (when $\beta_k > \alpha_k$) it has only been possible to show convergence for MVI (Hsieh et al., 2020). Instead of decreasing both stepsizes, Pethick et al. (2022) proposes a scheme that keeps the first stepsize constant,

$$\begin{aligned} \bar{z}^k &= z^k - \gamma \hat{F}(z^k, \xi_k) \quad \text{with} \quad \xi_k \sim \mathcal{P} \\ z^{k+1} &= z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k) \quad \text{with} \quad \bar{\xi}_k \sim \mathcal{P} \end{aligned} \tag{SEG+}$$

However, (SEG+) does not necessarily converge even in the monotone case as we illustrate in Figure 1. The non-convergence stems from the bias term introduced by the randomness of $\bar{z}^k$ in $\hat{F}(\bar{z}^k, \bar{\xi}_k)$. Intuitively, the role of $\bar{z}^k$ is to approximate the deterministic exploration step $\bar{\bar{z}}^k := z^k - \gamma F z^k$. While $\bar{z}^k$ is an unbiased estimate of $\bar{\bar{z}}^k$ this does not imply that $\hat{F}(\bar{z}^k, \bar{\xi}_k)$ is an unbiased estimate of $F(\bar{\bar{z}}^k)$. Unbiasedness only holds in special cases, such as when F is linear and $A \equiv 0$ for which we show convergence of (SEG+) in Section 5 under weak MVI. In the monotone case it suffice to take the exploration stepsize γ diminishing (Hsieh et al., 2020, Thm. 1), but this runs counter to the fixed stepsize requirement of weak MVI.

Instead we propose *bias-corrected stochastic extragradient+* (BCSEG+) in Algorithm 1. BCSEG+ adds a bias correction term of the previous operator evaluation using the current randomness ξ_k. This crucially allows us to keep the first stepsize fixed. We further generalize this scheme to constrained and regularized setting with Algorithm 2 by introducing the use of the resolvent, $(\mathrm{id} + \gamma A)^{-1}$.

5 Analysis of SEG+

(SEG+) does not converge in general even in the monotone case and even when considering the average iterate. Despite this failure, we can show convergence of (SEG+) under weak MVI in the special case where F is also linear and $A \equiv 0$.

Theorem 5.1. *Suppose that Assumptions I and II hold with $\Gamma = I$. Assume $Fz := Bz + v$ and choose $\alpha_k \in (0,1)$ and $\gamma \in (0, 1/L_F)$ such that $\rho \geq \gamma(\alpha_k - 1)/2$. Consider the sequence $(z^k)_{k \in \mathbb{N}}$ generated by (SEG+). Then for all $z^\star \in \mathcal{S}^\star$,*

$$\sum_{k=0}^{K} \frac{\alpha_k}{\sum_{j=0}^{K} \alpha_j} \mathbb{E}\|F z^k\|^2 \leq \frac{\|z^0 - z^\star\|^2 + \gamma^2(\gamma^2 L_F^2 + 1)\sigma_F^2 \sum_{j=0}^{K} \alpha_j^2}{\gamma^2(1 - \gamma^2 L_F^2) \sum_{j=0}^{K} \alpha_j}. \tag{5.1}$$

The underlying reason for this positive results is that $\hat{F}(\bar{z}^k, \bar{\xi}_k)$ is unbiased when F is linear. This no longer holds when either linearity of F is dropped or when the resolvent is introduced for $A \not\equiv 0$. An example of the failure is provided in Figure 1.

6 Analysis for unconstrained and smooth case

For simplicity we first consider the case where $A \equiv 0$. To mitigate the bias introduced in $F(\bar{z}^k, \bar{\xi}_k)$ for (SEG+), we propose Algorithm 1 which modifies the exploration step. The algorithm can be seen as a particular instance of the more general scheme treated in Section 7.

Theorem 6.1. *Suppose that Assumptions I to III hold with $\Gamma = I$. Suppose in addition that $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$ and $(\alpha_k)_{k\in\mathbb{N}} \subset (0,1)$ is a diminishing sequence such that*

$$2\gamma L_{\hat{F}} \sqrt{\alpha_0} + \left(1 + \left(\tfrac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^2 L_F^2\right)\gamma^2 L_{\hat{F}}^2\right)\alpha_0 \leq 1 + \tfrac{2\rho}{\gamma}. \tag{6.1}$$

Then, the following estimate holds for all $z^\star \in \mathcal{S}^\star$

$$\mathbb{E}[\|F(z^{k_\star})\|^2] \leq \frac{(1+\eta\gamma^2 L_F^2)\|z^0 - z^\star\|^2 + C\sigma_F^2\gamma^2 \sum_{j=0}^K \alpha_j^2}{\mu \sum_{j=0}^K \alpha_j} \tag{6.2}$$

where $C = 1 + 2\eta((\gamma^2 L_{\hat{F}}^2 + 1) + 2\alpha_0)$, $\eta = \frac{1}{2}\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^2 L_F^2 + \frac{1}{\gamma L_{\hat{F}}\sqrt{\alpha_0}}$, $\mu = \gamma^2(1-\gamma^2 L_F^2)/2$ and $k_\star$ is chosen from $\{0, 1, \ldots, K\}$ according to probability $\mathcal{P}[k_\star = k] = \frac{\alpha_k}{\sum_{j=0}^K \alpha_j}$.

Remark 6.2. As $\alpha_0 \to 0$, the requirement (6.1) reduces to $\rho > -\gamma/2$ as in the deterministic setting of Pethick et al. (2022). Letting $\alpha_k = \alpha_0/\sqrt{k+r}$ the rate becomes $\mathcal{O}(1/\sqrt{k})$, thus matching the rate of stochastic extragradient in the monotone case (see e.g. Juditsky et al. (2011)). □

The above result provides a rate for a *random iterate* as pioneered by Ghadimi & Lan (2013). Showing *last iterate* results even asymptotically is more challenging. Already in the monotone case, vanilla (SEG) (where $\beta_k = \alpha_k$) only has convergence guarantees for the average iterate (Juditsky et al., 2011). In fact, the scheme can cycle even in simple examples (Hsieh et al., 2020, Fig. 1).

Under the classical (but more restrictive) Robbins-Monro stepsize policy, it is possible to show almost sure convergence for the iterates generates by Algorithm 1. The following theorem demonstrates the result in the particular case of $\alpha_k = 1/k+r$. The more general statement is deferred to Appendix C.

Theorem 6.3 (almost sure convergence)**.** *Suppose that Assumptions I to III hold with $\Gamma = I$. Suppose $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$, $\alpha_k = \frac{1}{k+r}$ for any positive natural number r and*

$$(\gamma L_{\hat{F}} + 1)\alpha_k + \left(\tfrac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^4 L_F^4 \alpha_{k+1} + \gamma L_{\hat{F}}\right)(\alpha_{k+1} + 1)\alpha_{k+1} \leq 1 + \tfrac{2\rho}{\gamma}. \tag{6.3}$$

Then, the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 converges almost surely to some $z^\star \in \mathbf{zer}\, T$.

Remark 6.4. As $\alpha_k \to 0$ the condition on ρ reduces to $\rho > -\gamma/2$ like in the deterministic case. □

To make the results more accessible, both theorems have made particular choices of the free parameters from the proof, that ensures convergence for a given ρ and γ. However, since the parameters capture inherent tradeoffs, the choice above might not always provide the tightest rate. Thus, the more general statements of the theorems have been preserved in the appendix.

7 Analysis for constrained case

The result for the unconstrained smooth case can be extended when the resolvent is available. Algorithm 2 provides a direct generalization of the unconstrained Algorithm 1. The construction relies on approximating the deterministic algorithm proposed in Pethick et al. (2022), which iteratively projects onto a half-space which is guaranteed to contain the solutions. By defining $Hz = z - \gamma Fz$, the scheme can concisely be written as,

$$\begin{aligned} \bar{z}^k &= (I + \gamma A)^{-1}(Hz^k) \\ z^{k+1} &= z^k - \alpha_k(Hz^k - H\bar{z}^k), \end{aligned} \tag{CEG+}$$

for a particular adaptive choice of $\alpha_k \in (0,1)$. With a fair amount of hindsight we choose to replace Hz^k with the bias-corrected estimate h^k (as defined in Step 2.2 in Algorithm 2), such that the estimate is also reused in the second update.

Algorithm 2 (BCSEG+) Stochastic algorithm for problem (3.1)

REQUIRE $z^{-1} = z^0 \in \mathbb{R}^n$, $h^{-1} \in \mathbb{R}^n$, $\alpha_k \in (0,1)$, $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$
REPEAT for $k = 0, 1, \ldots$ until convergence
2.1: Sample $\xi_k \sim \mathcal{P}$
2.2: $h^k = (z^k - \gamma \hat{F}(z^k, \xi_k)) + (1-\alpha_k)\Big(h^{k-1} - (z^{k-1} - \gamma \hat{F}(z^{k-1}, \xi_k))\Big)$
2.3: $\bar{z}^k = (I + \gamma A)^{-1} h_k$
2.4: Sample $\bar{\xi}_k \sim \mathcal{P}$
2.5: $z^{k+1} = z^k - \alpha_k(h^k - \bar{z}^k + \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k))$
RETURN z^{k+1}

Theorem 7.1. *Suppose that Assumptions I and II hold with $\Gamma = I$. Moreover, suppose that $\alpha_k \in (0,1)$ and the following holds*

$$\mu := 1 + \frac{1+2\alpha_0\gamma^2 L_{\hat{F}}^2}{1-2\alpha_0\gamma^2 L_{\hat{F}}^2}\frac{2\rho}{\gamma} - \sqrt{\alpha_0} - \alpha_0 \frac{1+2\alpha_0\gamma^2 L_{\hat{F}}^2}{1-2\alpha_0\gamma^2 L_{\hat{F}}^2}(1 + 2\eta\gamma^2 L_{\hat{F}}^2) > 0 \tag{7.1}$$

where $\eta \geq \frac{2(1/(1-\gamma^2 L_{\hat{F}}^2)\sqrt{\alpha_0}) + (1+1/2\alpha_0\gamma^2 L_{\hat{F}}^2)(\alpha_0 - \frac{2\rho}{\gamma})}{1-2\alpha_0\gamma^2 L_{\hat{F}}^2}$. *Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2. Then, the following estimate holds for all $z^\star \in \mathcal{S}^\star$*

$$\mathbb{E}[\|H\bar{z}^{k_\star} - Hz^{k_\star}\|^2] \leq \frac{\mathbb{E}[\|z^0 - z^\star\|^2] + \eta\mathbb{E}[\|h^{-1} - Hz^{-1}\|^2] + C\gamma^2\sigma_F^2 \sum_{j=0}^K \alpha_j^2}{\mu \sum_{j=0}^K \alpha_j}$$

where $C = 1 + 2\eta(1+\gamma^2 L_{\hat{F}}^2) + 2\alpha_0(\frac{1}{(1-\gamma^2 L_{\hat{F}}^2)\sqrt{\alpha_0}} + (\frac{1}{2} + \frac{1}{4\alpha_0\gamma^2 L_{\hat{F}}^2})(\alpha_0 - \frac{2\rho}{\gamma} + \alpha_0 2\gamma^2 L_{\hat{F}}^2\eta))$ *and $k_\star$ is chosen from $\{0, 1, \ldots, K\}$ according to probability $\mathcal{P}[k_\star = k] = \frac{\alpha_k}{\sum_{j=0}^K \alpha_j}$.*

Remark 3. The condition on ρ in (7.1) reduces to $\rho > -\gamma/2$ when $\alpha_0 \to 0$ as in the deterministic case. Note that Algorithm 1 and Algorithm 2 coincide when $A \equiv 0$, so Theorem 7.1 also applies to Algorithm 1 in the unconstrained case.

8 ASYMMETRIC PRECONDITIONING

The family of stochastic algorithms which converges under weak MVI can be expanded beyond Algorithm 2, by considering a nonlinear asymmetric preconditioner for $u, z \in \mathbb{R}^n$,

$$P_u(z) := \Gamma^{-1} z + Q_u(z), \tag{8.1}$$

where $Q_u(z)$ is the asymmetric component to be defined. With $P_u(z)$ at hand, we can generalize the deterministic (CEG+) scheme from Section 7 as follows

$$\text{find } \bar{z}^k \text{ such that} \quad H_{z^k}(z^k) \in P_{z^k}(\bar{z}^k) + A(\bar{z}^k) \tag{8.2a}$$

$$\text{update} \quad z^{k+1} = z^k + \alpha\Gamma\Big(H_{z^k}(\bar{z}^k) - H_{z^k}(z^k)\Big) \tag{8.2b}$$

where $H_u(z) := P_u(z) - F(z)$. The template (8.2) produces a different algorithm depending on the choice of $Q_u(z)$. By defining the stochastic variant of (8.1),

$$\hat{P}_u(z, \xi) := \Gamma^{-1} z + \hat{Q}_u(z, \xi), \tag{8.3}$$

we can present the stochastic counterpart, which introduces a carefully chosen bias-correction term.

$$\text{compute} \quad h^k = \hat{P}_{z^k}(z^k, \xi_k) - \hat{F}(z^k, \xi_k) + (1-\alpha_k)\Big(h^{k-1} - \hat{P}_{z^{k-1}}(z^{k-1}, \xi_k) + \hat{F}(z^{k-1}, \xi_k) \tag{8.4a}$$
$$- \hat{Q}_{z^{k-1}}(\bar{z}^{k-1}, \xi'_{k-1}) + \hat{Q}_{z^{k-1}}(\bar{z}^{k-1}, \xi'_k)\Big) \qquad \text{with} \quad \xi_k, \xi'_k \sim \mathcal{P}$$

$$\text{find } \bar{z}^k \text{ such that} \quad h^k \in \hat{P}_{z^k}(\bar{z}^k, \xi'_k) + A\bar{z}^k \tag{8.4b}$$

$$\text{update} \quad z^{k+1} = z^k + \alpha_k\Gamma\Big(\hat{P}_{z^k}(\bar{z}^k, \bar{\xi}_k) - \hat{F}(\bar{z}^k, \bar{\xi}_k) - h^k\Big) \qquad \text{with} \quad \bar{\xi}_k \sim \mathcal{P} \tag{8.4c}$$

Remark 4. The two additional terms in the bias-correction term is due to the interesting interplay between weak MVI and stochastic feedback, which forces a change of variables (see Appendix E.1).

To make a concrete choice of $\hat{Q}_u(z, \xi)$ we will consider a minimax problem as a motivating example.

Algorithm 3 Nonlinearly preconditioned primal dual hybrid gradient (NP-PDHG) for solving (8.5)

REQUIRE $z^{-1} = z^0 = (x^0, y^0)$ with $x^0, x^{-1}, \hat{x}^{-1}, \bar{x}^{-1} \in \mathbb{R}^n$, $y^0, y^{-1} \in \mathbb{R}^r$, $\theta \in [0, \infty)$, $\Gamma_1 \succ 0$, $\Gamma_2 \succ 0$
REPEAT for $k = 0, 1, \ldots$ until convergence
3.1: $\xi_k \sim \mathcal{P}$
3.2: $\hat{x}^k = x^k - \Gamma_1 \nabla_x \hat{\varphi}(z^k, \xi_k) + (1 - \alpha_k)(\hat{x}^{k-1} - x^{k-1} + \Gamma_1 \nabla_x \hat{\varphi}(x^{k-1}, y^{k-1}, \xi_k))$
3.3: $\bar{x}^k = \mathbf{prox}_f^{\Gamma_1^{-1}}(\hat{x}^k)$
3.4: $\xi'_k \sim \mathcal{P}$
3.5: $\hat{y}^k = y^k + \Gamma_2(\theta \nabla_y \hat{\varphi}(\bar{x}^k, y^k, \xi'_k) + (1 - \theta)\nabla_y \hat{\varphi}(z^k, \xi_k))$
3.6: $+(1 - \alpha_k)\Big(\hat{y}^{k-1} - y^{k-1} - \Gamma_2(\theta \nabla_y \hat{\varphi}(\bar{x}^{k-1}, y^{k-1}, \xi'_k) + (1 - \theta)\nabla_y \hat{\varphi}(z^{k-1}, \xi_k))\Big)$
3.7: $\bar{y}^k = \mathbf{prox}_g^{\Gamma_2^{-1}}(\hat{y}^k)$
3.8: $\bar{\xi}_k \sim \mathcal{P}$
3.9: $x^{k+1} = x^k + \alpha_k\Big(\bar{x}^k - \hat{x}^k - \Gamma_1 \nabla_x \hat{\varphi}(\bar{z}^k, \bar{\xi}_k)\Big)$
3.10: $y^{k+1} = y^k + \alpha_k\Big(\bar{y}^k - \hat{y}^k + \Gamma_2 \nabla_y \hat{\varphi}(\bar{z}^k, \bar{\xi}_k)\Big)$
RETURN $z^{k+1} = (x^{k+1}, y^{k+1})$

8.1 NONLINEARLY PRECONDITIONED PDHG

We consider the problem of

$$\underset{x \in \mathbb{R}^n}{\textbf{minimize}}\ \underset{y \in \mathbb{R}^r}{\textbf{maximize}} \quad f(x) + \varphi(x, y) - g(y). \tag{8.5}$$

where $\varphi(x, y) := \mathbb{E}_\xi[\hat{\varphi}(x, y, \xi)]$. The first order optimality conditions may be written as the inclusion

$$0 \in Tz := Az + Fz, \quad \text{where} \quad A = (\partial f, \partial g), \quad F(z) = (\nabla_x \varphi(z), -\nabla_y \varphi(z)). \tag{8.6}$$

Assumption IV. *In problem* (8.5)*,*

A1 *f, g are proper lsc convex*

A2 *$\varphi : \mathbb{R}^{n+r} \to \mathbb{R}$ is continuously differentiable and for some symmetric positive definite matrices $D_{xx}, D_{xy}, D_{yx}, D_{yy}$, the following holds for all $z = (x, y), z' = (x', y') \in \mathbb{R}^{n+r}$*

$$\|\nabla_x \varphi(z') - \nabla_x \varphi(z)\|_{\Gamma_1}^2 \le \|x' - x\|_{D_{xx}}^2 + \|y' - y\|_{D_{xy}}^2,$$
$$\|\nabla_y \varphi(z') - \theta \nabla_y \varphi(x', y) - (1 - \theta)\nabla_y \varphi(z)\|_{\Gamma_2}^2 \le \|x' - x\|_{D_{yx}}^2 + \|y' - y\|_{D_{yy}}^2.$$

A3 *(stepsize condition)*

$$D_{xx} + D_{yx} \succeq \Gamma_1, \qquad D_{xy} + D_{yy} \succeq \Gamma_2. \tag{8.7}$$

A4 *$\hat{\varphi}(\cdot, \xi) : \mathbb{R}^{n+r} \to \mathbb{R}$ is continuously differentiable and for some symmetric positive definite matrices $D_{\widehat{xz}}, D_{\widehat{yz}}, D_{\widehat{yx}}, D_{\widehat{yy}}$, the following holds for all $z = (x, y), z' = (x', y') \in \mathbb{R}^{n+r}$ and $v, v' \in \mathbb{R}^n$*

$$\mathbb{E}_\xi\Big[\|\nabla_x \hat{\varphi}(z', \xi) - \nabla_x \hat{\varphi}(z, \xi)\|_{\Gamma_1}^2\Big] \le \|z' - z\|_{D_{\widehat{xz}}}^2$$
$$\mathbb{E}_{\xi, \xi'}\Big[\|(1 - \theta)(\nabla_y \hat{\varphi}(z, \xi) - \nabla_y \hat{\varphi}(z', \xi)) - \theta(\nabla_y \hat{\varphi}(v', y', \xi') - \nabla_y \hat{\varphi}(v, y, \xi'))\|_{\Gamma_2}^2\Big]$$
$$\le \|z' - z\|_{D_{\widehat{yz}}}^2 + \|\bar{x}' - \bar{x}\|_{D_{\widehat{yx}}}^2 + \|y' - y\|_{D_{\widehat{yy}}}^2.$$

Remark 8.1. In Algorithm 3 the choice of $\theta \in [0, \infty)$ leads to different algorithmic oracles and underlying assumptions in terms of Lipschitz continuity in requirement IV.A2.

(i) If $\theta = 0$ then the first two steps may be computed in parallel and we recover Algorithm 2. Moreover, to ensure requirement IV.A2 in this case it suffices to assume

$$\|\nabla_x \varphi(z') - \nabla_x \varphi(z)\| \le L_x \|z' - z\|, \quad \|\nabla_y \varphi(z') - \nabla_y \varphi(z)\| \le L_y \|z' - z\|.$$

(ii) Taking $\theta = 1$ leads to Gauss-Seidel updates and Algorithm 3 captures a nonlinear extension of the celebrated primal dual hybrid gradient (PDHG) (Chambolle & Pock, 2011) with sufficient Lipschitz continuity assumptions for some $L_x, L_y \in [0, \infty)$,

$$\|\nabla_x \varphi(z') - \nabla_x \varphi(z)\| \le L_x \|z' - z\|, \quad \|\nabla_y \varphi(z') - \nabla_y \varphi(x', y)\| \le L_y \|y' - y\|.$$

□

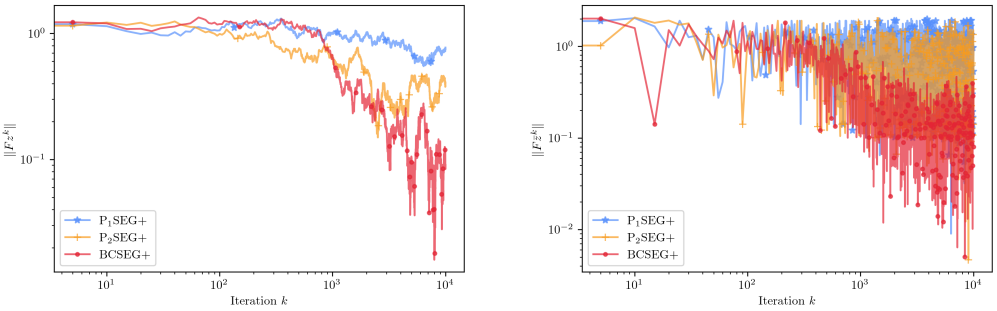

Figure 1: *Monotone constrained case illustrating the failure mode of* (SEG+)*. The objective is bilinear* $\phi(x, y) = (x - 0.9) \cdot (y - 0.9)$ *under box constraints* $\|(x, y)\|_\infty \leq 1$*. The unique stationary point* $(x^\star, y^\star) = (0.9, 0.9)$ *lies in the interior, so even* $\|Fz\|$ *can be driven to zero. Despite the simplicity of the problem neither projected variants of* (SEG+) *converges (see Appendix G.2 for algorithmic details).*

Algorithm 3 is an application of (8.4) applied for solving (8.6). In order to cast the algorithm as an instance of the template algorithm (8.4), we choose the positive definite stepsize matrix as

$$\Gamma = \mathbf{blkdiag}(\Gamma_1, \Gamma_2), \quad \Gamma_1 \succ 0,\ \Gamma_2 \succ 0.$$

and the nonlinear part of the preconditioner as

$$\hat{Q}_u(\bar{z}, \xi) := \Big(0, -\theta \nabla_y \hat{\varphi}(\bar{x}, y, \xi)\Big), \quad \text{and} \quad Q_u(\bar{z}) := \Big(0, -\theta \nabla_y \varphi(\bar{x}, y)\Big) \tag{8.8}$$

where $u = (x, y)$ and $\bar{z} = (\bar{x}, \bar{y})$. Recall $H_u(z) := P_u(z) - F(z)$ and define $S_u(z; \bar{z}) := H_u(z) - Q_u(\bar{z})$. The convergence in Theorem 8.2 depends on the distance between the initial estimate $\Gamma^{-1}\hat{z}^{-1}$ with $\hat{z}^{-1} = (\hat{x}^{-1}, \hat{y}^{-1})$ and the deterministic $S_{z^{-1}}(z^{-1}; \bar{z}^{-1})$.

Theorem 8.2. *Suppose that Assumption I(ii) to II and IV hold. Moreover, suppose that* $\alpha_k \in (0, 1)$*,* $\theta \in [0, 1]$ *and the following holds,*

$$\mu := 1 + (1 + \tfrac{2\alpha_0 d}{1-\alpha_0(4\hat{c}_2+d)})\tfrac{2\rho}{\bar{\gamma}} - \sqrt{\alpha_0} - \alpha_0(1 + \tfrac{2\alpha_0 d}{1-\alpha_0(4\hat{c}_2+d)})(1 + \hat{c}_4\eta) > 0 \tag{8.9}$$

where $\bar{\gamma}$ *denotes the smallest eigenvalue of* Γ*,* $\eta \geq 2(\frac{1}{(1-L_M^2)\sqrt{\alpha_0}} + \frac{1+\alpha_0(d-4\hat{c}_2)}{2\alpha_0 d}(\alpha_0 - \frac{2\rho}{\bar{\gamma}}))/(1-\alpha_0(4\hat{c}_2+d))$ *and*

$$\hat{c}_1 = \|\Gamma(D_{\widehat{x}\widetilde{z}} + D_{\widetilde{y}\widetilde{z}}) + \Gamma_2 D_{\widehat{y}\widetilde{y}}\|, \quad \hat{c}_2 = \|\Gamma_1 D_{\widehat{y}\widetilde{x}}\|, \quad \hat{c}_3 = \|D_{\widehat{x}\widetilde{z}}\Gamma\|, \quad \hat{c}_4 = 2(\hat{c}_1 + 2\hat{c}_2(1 + \hat{c}_3))$$

$$d = 2(1 + 4\hat{c}_2\alpha_0^2)(\hat{c}_1 + 2\hat{c}_2(1 + \hat{c}_3)), \quad L_M^2 = \mathbf{max}\,\{\|D_{xx}\Gamma_1\| + \|D_{xy}\Gamma_2\|, \|D_{yx}\Gamma_1\| + \|D_{yy}\Gamma_2\|\}.$$

Consider the sequence $(z^k)_{k\in\mathbb{N}}$ *generated by Algorithm 3. Then, the following holds for all* $z^\star \in \mathcal{S}^\star$

$$\mathbb{E}[\|H_{z^{k_\star}}(\bar{z}^{k_\star}) - H_{z^{k_\star}}(z^{k_\star})\|_\Gamma^2] \leq \frac{\mathbb{E}[\|z^0 - z^\star\|_{\Gamma^{-1}}^2] + \eta\mathbb{E}[\|\Gamma^{-1}\hat{z}^{-1} - S_{z^{-1}}(z^{-1}; \bar{z}^{-1})\|_\Gamma^2] + C\sigma_F^2 \sum_{j=0}^K \alpha_j^2}{\mu \sum_{j=0}^K \alpha_j}$$

where $k_\star$ *is chosen from* $\{0, 1, \ldots, K\}$ *according to probability* $\mathcal{P}[k_\star = k] = \frac{\alpha_k}{\sum_{j=0}^K \alpha_j}$ *and*

$$C = 2(\eta + \alpha_0(\tfrac{1}{(1-L_M^2)\sqrt{\alpha_0}} + \tfrac{1+\alpha_0(d-4\hat{c}_2)}{2\alpha_0 d}(\alpha_0 - 2\tfrac{\rho}{\bar{\gamma}} + \alpha_0\hat{c}_4\eta)))(1 + 2\hat{c}_2) + 1 + \hat{c}_4\eta.$$

Remark 5. When $\alpha_0 \to 0$ the conditions in (8.2) reduces to $1 + \frac{2\rho}{\bar{\gamma}} > 0$ as in the deterministic case.

For $\theta = 0$ Algorithm 3 reduces to Algorithm 2. With this choice Theorem 8.2 simplifies, since the constant $\hat{c}_2 = 0$, and we recover the convergence result of Theorem 7.1.

9 Experiments

We compare BCSEG+ against (SEG) and (EG+) in both an unconstrained setting and a constrained setting introduced in Pethick et al. (2022). In the unconstrained example we control all problem constant and set $\rho = -1/10L_F$, while the constrained example is a specific minimax problem where $\rho > -1/2L_F$ holds within the constraint set for a Lipschitz constant L_F restricted to the same constraint set. To simulate a stochastic setting in both examples, we consider additive Gaussian noise, i.e.

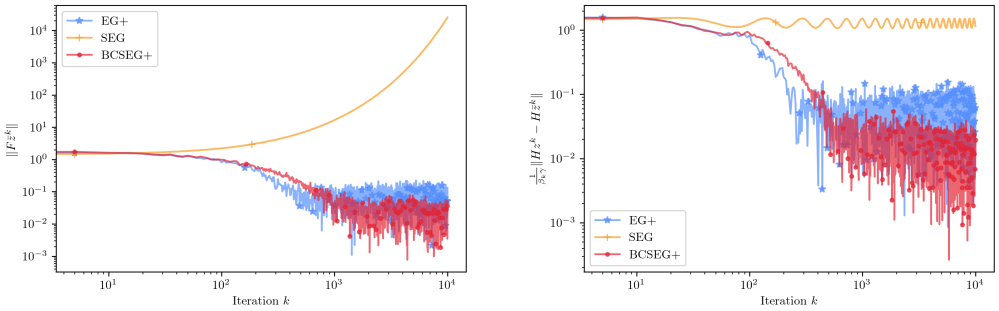

Figure 2: *Comparison of methods in the unconstrained setting of Example 2 (left) and the constrained setting of Example 3 (right). Notice that, except for BCSEG+, all methods fails to converge in these examples.*

$\hat{F}(z, \xi) = Fz + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2 I)$. In the experiments we choose $\sigma = 0.1$. Further details can be found in the supplementary material including constraint extensions of baseline algorithms.

The results are show in Figure 2. The sequence generated by (SEG) diverges for the unconstrained problem and cycles in the constrained problem. In comparison (EG+) gets closer to a solution in both problems but fails to converge due to the non-diminishing stepsize, while BCSEG+ converges for both example.

10 Conclusion

This paper shows that nonconvex-nonconcave problems characterize by the weak Minty variational inequality can be solved efficiently even when only *stochastic* gradients are available. The approach crucially avoids increasing batch sizes by instead introducing a bias-correction term. We show that convergence is possible for the same range of problem constant $\rho \in (-\gamma/2, \infty)$ as in the deterministic case. Rates are established for a *random iterate*, which matches those of stochastic extragradient in the monotone case, and the result is complemented with almost sure convergence, thus providing asymptotic convergence for the *last iterate*. We show that the idea extends to a family of extragradient-type methods which includes a nonlinear extension of the celebrated primal dual hybrid gradient (PDHG) algorithm. For future work it is interesting to see if the rate can be improved by considering accelerated methods such as Halpern iterations.

References

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Heinz H Bauschke, Walaa M Moursi, and Xianfu Wang. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189(1):55–74, 2021.

Dimitri P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

Axel Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. *arXiv preprint arXiv:2201.12247*, 2022.

Xufeng Cai, Chaobing Song, Cristóbal Guzmán, and Jelena Diakonikolas. A stochastic halpern iteration with variance reduction for stochastic monotone inclusion problems. *arXiv preprint arXiv:2203.09436*, 2022.

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

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Patrick L Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1466–1478, 2021.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pp. 7865–7901. PMLR, 2022.

M Hirsch and S Vavasis. Exponential lower bounds for finding Brouwer fixed points. In *Proceedings of the 28th Symposium on Foundations of Computer Science*, pp. 401–410, 1987.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *arXiv preprint arXiv:2003.10162*, 2020.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2106.02326*, 2021.

Chris Junchi Li, Yaodong Yu, Nicolas Loizou, Gauthier Gidel, Yi Ma, Nicolas Le Roux, and Michael Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pp. 9793–9826. PMLR, 2022.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.

Lam M Nguyen, Marten van Dijk, Dzung T Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R Kalagnanam. Finite-sum smooth optimization with sarah. *arXiv preprint arXiv:1901.07648*, 2019.

Christos H Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences*, 48(3):498–532, 1994.

Thomas Pethick, Puya Latafat, Panagiotis Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

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

Appendix

Table of Contents

A Prelude **13**

B Proof for SEG+ **13**

Proof of Theorem 5.1 13

C Proof for smooth unconstrained case **14**

Proof of Theorem C.2 14

Proof of Theorem 6.1 16

Proof of Theorem C.3 17

Proof of Theorem 6.3 18

D Proof for constrained case **18**

Proof of Theorem D.1 19

Proof of Theorem 7.1 21

E Proof for nonlinearly preconditioned PDHG through asymmetric preconditioner **21**

Proof of Theorem E.1 22

Proof of Theorem 8.2 26

E.1 Explanation of bias-correction term 26

F Negative weak Minty variational inequality **27**

G Experiments **27**

G.1 Synthetic example 27

G.2 Additional algorithmic details 28

H Comparison with variance reduction **28**

Table 1: *Overview of the results. The second row is obtained as special cases of the first row.*

	Unconstrained & smooth ($A \equiv 0$)		Constrained ($A \not\equiv 0$)	
	Random iterate	Last iterate	BCSEG+	NP-PDHG
Appendix	Theorem C.2	Theorem C.3	Theorem D.1	Theorem E.1
	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
Main paper	Theorem 6.1	Theorem 6.3	Theorem 7.1	Theorem 8.2

A Prelude

For the unconstrained and smooth setting Appendix B treats convergences of (SEG+) for the restricted case where F is linear. Appendix C shows both random iterate results and almost sure convergence of Algorithm 1. Theorems 6.1 and 6.3 in the main body are implied by the more general results in this section, which preserves certain free parameters and more general stepsize requirements. Appendices D and E moves beyond the unconstrained and smooth case by showing convergence for instances of the template scheme (8.2). The analysis of Algorithm 3 in Appendix E applies to Algorithm 2, but for completeness we establish convergence for general F separately in Appendix D. The relationship between the theorems are presented in Table 1.

B Proof for SEG+

Proof of Theorem 5.1. Following (Hsieh et al., 2020) closely, define the reference state $\tilde{\bar{z}}^k := z^k - \gamma F z^k$ to be the exploration step using the *deterministic* operator and denote the second stepsize as $\eta_k := \alpha_k \gamma$. We will let ζ denote the additive noise term, i.e. $\hat{F}(z, \xi) := F(z) + \zeta$. Expanding the distance to solution,

$$\begin{aligned}
\|z^{k+1} - z^\star\|^2 &= \|z^k - \eta_k \hat{F}(\bar{z}^k, \bar{\xi}_k) - z^\star\|^2 \\
&= \|z^k - z^\star\|^2 - 2\eta_k \langle \hat{F}(\bar{z}^k, \bar{\xi}_k), z^k - z^\star \rangle + \eta_k^2 \|\hat{F}(\bar{z}^k, \bar{\xi}_k)\|^2 \\
&= \|z^k - z^\star\|^2 - 2\eta_k \langle \hat{F}(\bar{z}^k, \bar{\xi}_k), \tilde{\bar{z}}^k - z^\star \rangle - 2\gamma\eta_k \langle \hat{F}(\bar{z}^k, \bar{\xi}_k), F(z^k) \rangle + \eta_k^2 \|\hat{F}(\bar{z}^k, \bar{\xi}_k)\|^2.
\end{aligned} \tag{B.1}$$

Recall that the operator is assumed to be linear $Fz = Bz + v$ in which case we have,

$$\begin{aligned}
\hat{F}(\bar{z}^k, \bar{\xi}_k) &= B\bar{z}^k + v + \bar{\zeta}_k \\
&= B(z^k - \gamma \hat{F}(z^k, \xi_k)) + v + \bar{\zeta}_k \\
&= B(z^k - \gamma B z^k - \gamma v - \gamma \zeta_k) + v + \bar{\zeta}_k \\
&= B(z^k - \gamma(Bz^k + v)) + v - \gamma B \zeta_k + \bar{\zeta}_k \\
&= F(\tilde{\bar{z}}^k) - \gamma B \zeta_k + \bar{\zeta}_k.
\end{aligned} \tag{B.2}$$

The two latter term are zero in expectation due to the unbiasedness from Assumption II*(i)*, which lets us write the terms on the RHS of (B.1) as,

$$-\mathbb{E}_k \langle \hat{F}(\bar{z}^k, \bar{\xi}_k), \tilde{\bar{z}}^k - z^\star \rangle = -\langle F(\tilde{\bar{z}}^k), \tilde{\bar{z}}^k - z^\star \rangle \tag{B.3}$$

$$-\mathbb{E}_k \langle \hat{F}(\bar{z}^k, \bar{\xi}_k), F(z^k) \rangle = -\langle F(\tilde{\bar{z}}^k), F(z^k) \rangle \tag{B.4}$$

$$\mathbb{E}_k \|\hat{F}(\bar{z}^k, \bar{\xi}_k)\|^2 = \|F(\tilde{\bar{z}}^k)\|^2 + \mathbb{E}_k \|\gamma B \zeta_k\|^2 + \mathbb{E}_k \|\bar{\zeta}_k\|^2. \tag{B.5}$$

We can bound (B.3) directly through the weak MVI in Assumption I*(iii)* which might still be positive,

$$-\langle F(\tilde{\bar{z}}^k), \tilde{\bar{z}}^k - z^\star \rangle \le -\rho \|F(\tilde{\bar{z}}^k)\|^2. \tag{B.6}$$

For the latter two terms of (B.5) we have

$$\mathbb{E}_k \|\gamma B \zeta_k\|^2 + \mathbb{E}_k \|\bar{\zeta}_k\|^2 = \gamma^2 \mathbb{E}_k \|F(\zeta_k)\|^2 + \mathbb{E}_k \|\bar{\zeta}_k\|^2 \le (\gamma^2 L_F^2 + 1)\sigma_F^2, \tag{B.7}$$

where the last inequality follows from Lipschitz in Assumption I*(i)* and bounded variance in Assumption II*(ii)*.

Combining everything into (B.1) we are left with

$$\mathbb{E}_k \|z^{k+1} - z^\star\|^2 \le \|z^k - z^\star\|^2 + \eta_k^2(\gamma^2 L_F^2 + 1)\sigma_F^2 - 2\gamma\eta_k \langle F(\tilde{\bar{z}}^k), F(z^k) \rangle + (\eta_k^2 - 2\eta_k \rho)\|F(\tilde{\bar{z}}^k)\|^2 \tag{B.8}$$

By assuming the stepsize condition, $\rho \geq (\eta_k - \gamma)/2$, we have $\eta_k^2 - 2\eta_k\rho \leq \gamma\eta_k$. This allows us to complete the square,

$$\begin{aligned}-2\gamma\eta_k\langle F(\tilde{z}^k), F(z^k)\rangle + (\eta_k^2 - 2\eta_k\rho)\|F(\tilde{z}^k)\|^2 &\leq -2\gamma\eta_k\langle F(\tilde{z}^k), F(z^k)\rangle + \gamma\eta_k\|F(\tilde{z}^k)\|^2 \\ &= \gamma\eta_k(\|F(z^k) - F(\tilde{z}^k)\|^2 - \|F(z^k)\|^2) \\ &\leq \gamma\eta_k(\gamma^2 L_F^2 - 1)\|F(z^k)\|^2,\end{aligned} \tag{B.9}$$

where the last inequality follows from Lipschitzness of F and the definition of the update rule. Plugging into (B.8) we are left with

$$\mathbb{E}_k\|z^{k+1} - z^\star\|^2 \leq \|z^k - z^\star\|^2 + \eta_k^2(\gamma^2 L_F^2 + 1)\sigma_F^2 + \gamma\eta_k(\gamma^2 L_F^2 - 1)\|F(z^k)\|^2. \tag{B.10}$$

The result is obtained by total expectation and summing. □

C Proof for smooth unconstrained case

Lemma C.1. *Consider the recurrent relation $B_{k+1} = \xi_k B_k + d_k$ such that $\xi_k > 0$ for all $k \geq 0$. Then*

$$B_{k+1} = (\Pi_{p=0}^k \xi_p)\left(B_0 + \sum_{\ell=0}^k \frac{d_\ell}{\Pi_{p=0}^\ell \xi_p}\right).$$

Assumption V. $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$ *and*

$$\mu := \gamma^2(1 - \gamma^2 L_F^2(1 + b^{-1})) > 0 \tag{C.1}$$

Theorem C.2. *Suppose that Assumptions I to III hold with $\Gamma = I$. Suppose in addition that Assumption V holds and that $(\alpha_k)_{k\in\mathbb{N}} \subset (0, 1)$ is a diminishing sequence such that*

$$2\gamma L_{\hat{F}}\sqrt{\alpha_0} + \left(1 + ((b+1)\gamma^2 L_F^2)\gamma^2 L_{\hat{F}}^2\right)\alpha_0 \leq 1 + \tfrac{2\rho}{\gamma}. \tag{C.2}$$

Then, the following estimate holds

$$\sum_{k=0}^K \frac{\alpha_k}{\sum_{j=0}^K \alpha_j}\mathbb{E}[\|F(z^k)\|^2] \leq \frac{\|z^0 - z^\star\|^2 + \eta\gamma^2\|F(z^0)\|^2 + C\sigma_F^2\gamma^2\sum_{j=0}^K \alpha_j^2}{\mu\sum_{j=0}^K \alpha_j}, \tag{C.3}$$

where $C = 1 + 2\eta((\gamma^2 L_{\hat{F}}^2 + 1) + 2\alpha_0)$ and $\eta = \frac{1}{2}(b+1)\gamma^2 L_F^2 + \frac{1}{\gamma L_{\hat{F}}\sqrt{\alpha_0}}$.

Proof of Theorem C.2. The proof relies on establishing a (stochastic) descent property on the following function

$$\mathcal{U}_{k+1} := \|z^{k+1} - z^\star\|^2 + A_{k+1}\|u^k\|^2 + B_{k+1}\|z^{k+1} - z^k\|^2.$$

where $u^k := \bar{z}^k - z^k + \gamma F(z^k)$, and $(A_k)_{k\in\mathbb{N}}$, $(B_k)_{k\in\mathbb{N}}$ are positive scalar parameters to be identified. We proceed to consider each term individually.

Let us begin by quantifying how well $\bar{z}^k$ estimates $z^k - \gamma F(z^k)$.

$$u^k = \bar{z}^k - z^k + \gamma F(z^k) = \gamma F(z^k) - \gamma\hat{F}(z^k, \xi_k) + (1 - \alpha_k)(\bar{z}^{k-1} - z^{k-1} + \gamma\hat{F}(z^{k-1}, \xi_k)).$$

Therefore,

$$\begin{aligned}\|u^k\|^2 = {}& \|\gamma F(z^k) - \gamma\hat{F}(z^k, \xi_k) + (1 - \alpha_k)(\gamma\hat{F}(z^{k-1}, \xi_k) - \gamma F(z^{k-1}))\|^2 + (1 - \alpha_k)^2\|u^{k-1}\|^2 \\ &+ 2(1 - \alpha_k)\langle \bar{z}^{k-1} - z^{k-1} + \gamma\hat{F}(z^{k-1}), \gamma F(z^k) - \gamma\hat{F}(z^k, \xi_k) + (1 - \alpha_k)(\gamma\hat{F}(z^{k-1}, \xi_k) - \gamma F(z^{k-1}))\rangle.\end{aligned}$$

Conditioned on $\mathcal{F}_k$, in the inner product the left term is known and the right term has an expectation that equals zero. Therefore, we obtain

$$\begin{aligned}\mathbb{E}[\|u^k\|^2|\mathcal{F}_k] &= \mathbb{E}[\|(1-\alpha_k)\left(\gamma F(z^k) - \gamma\hat{F}(z^k,\xi_k) + \gamma\hat{F}(z^{k-1},\xi_k) - \gamma F(z^{k-1})\right) + \alpha_k\left(\gamma F(z^k) - \gamma\hat{F}(z^k,\xi_k)\right)\|^2|\mathcal{F}_k] \\ &\quad + (1-\alpha_k)^2\|u^{k-1}\|^2 \\ &\leq (1-\alpha_k)^2\|u^{k-1}\|^2 + 2(1-\alpha_k)^2\gamma^2\mathbb{E}[\|\hat{F}(z^k,\xi_k) - \hat{F}(z^{k-1},\xi_k)\|^2|\mathcal{F}_k] \\ &\quad + 2\alpha_k^2\gamma^2\mathbb{E}[\|F(z^k) - \hat{F}(z^k,\xi_k)\|^2|\mathcal{F}_k] \\ &\leq (1-\alpha_k)^2\|u^{k-1}\|^2 + 2(1-\alpha_k)^2\gamma^2 L_{\hat{F}}^2\|z^k - z^{k-1}\|^2 + 2\alpha_k^2\gamma^2\sigma_F^2\end{aligned} \tag{C.4}$$

where in the first inequality we used Young inequality and the fact that the second moment is larger than the variance, and Assumptions II*(ii)* and III were used in the second inequality.

By step 1.2, the equality

$$\|z^{k+1} - z^\star\|^2 = \|z^k - z^\star\|^2 - 2\alpha_k\gamma\langle \hat{F}(\bar{z}^k, \bar{\xi}_k), z^k - z^\star\rangle + \alpha_k^2\gamma^2\|\hat{F}(\bar{z}^k, \bar{\xi}_k)\|^2, \tag{C.5}$$

holds. The inner product in (C.5) can be upper bounded using Young inequalities with positive parameters ε_k, $k \geq 0$, and b as follows.

$$\begin{aligned}
\mathbb{E}[\langle -\gamma\hat{F}(\bar{z}^k, \bar{\xi}_k), z^k - z^\star\rangle \mid \bar{\mathcal{F}}_k] =& -\gamma\langle F(\bar{z}^k), z^k - \bar{z}^k\rangle - \gamma\langle F(\bar{z}^k), \bar{z}^k - z^\star\rangle \\
=& -\gamma^2\langle F(\bar{z}^k), F(z^k)\rangle + \gamma\langle F(\bar{z}^k), \bar{z}^k - z^k + \gamma F(z^k)\rangle - \gamma\langle F(\bar{z}^k), \bar{z}^k - z^\star\rangle \\
\leq& \gamma^2\Big(\frac{1}{2}\|F(\bar{z}^k) - F(z^k)\|^2 - \frac{1}{2}\|F(\bar{z}^k)\|^2 - \frac{1}{2}\|F(z^k)\|^2\Big) + \frac{\gamma^2\varepsilon_k}{2}\|F(\bar{z}^k)\|^2 \\
&+ \frac{1}{2\varepsilon_k}\|\bar{z}^k - z^k + \gamma F(z^k)\|^2 - \gamma\rho\|F(\bar{z}^k)\|^2 \\
\leq& \gamma^2 L_F^2\frac{1+b}{2}\|u^k\|^2 + \frac{1+b^{-1}}{2}\gamma^4 L_F^2\|F(z^k)\|^2 - \frac{\gamma^2}{2}\|F(\bar{z}^k)\|^2 \\
&- \frac{\gamma^2}{2}\|F(z^k)\|^2 - \frac{\gamma^2\varepsilon_k}{2}\|F(\bar{z}^k)\|^2 + \frac{1}{2\varepsilon_k}\|u^k\|^2 - \gamma\rho\|F(\bar{z}^k)\|^2 \\
=& (\gamma^2 L_F^2\frac{1+b}{2} + \frac{1}{2\varepsilon_k})\|u^k\|^2 + \frac{\gamma^2(\gamma^2 L_F^2(1+b^{-1}) - 1)}{2}\|F(z^k)\|^2 \\
&+ (\frac{\gamma^2(\varepsilon_k - 1)}{2} - \gamma\rho)\|F(\bar{z}^k)\|^2.
\end{aligned} \tag{C.6}$$

Conditioning (C.6) with $\mathbb{E}[\cdot \mid \mathcal{F}_k] = \mathbb{E}[\mathbb{E}[\cdot \mid \bar{\mathcal{F}}_k] \mid \mathcal{F}_k]$, since $\mathcal{F}_k \subset \bar{\mathcal{F}}_k$, yields

$$\begin{aligned}
2\mathbb{E}[\langle -\gamma\hat{F}(\bar{z}^k, \bar{\xi}_k), z^k - z^\star\rangle \mid \mathcal{F}_k] \leq& (\gamma^2 L_F^2(1+b) + \frac{1}{\varepsilon_k})\mathbb{E}[\|u^k\|^2 \mid \mathcal{F}_k] - \mu\|F(z^k)\|^2 \\
&+ (\gamma^2(\varepsilon_k - 1) - 2\gamma\rho)\mathbb{E}\Big[\|F(\bar{z}^k)\|^2 \mid \mathcal{F}_k\Big],
\end{aligned} \tag{C.7}$$

where μ was defined in (C.1).

The third term in (C.5) is bounded through Assumption II*(ii)* by

$$\mathbb{E}[\|\hat{F}(\bar{z}^k, \xi_k)\|^2 \mid \bar{\mathcal{F}}_k] \leq \|F(\bar{z}^k)\|^2 + \sigma_F^2,$$

which in turn implies

$$\mathbb{E}\Big[\|z^{k+1} - z^k\|^2 \mid \mathcal{F}_k\Big] = \alpha_k^2\gamma^2\mathbb{E}\Big[\|\hat{F}(\bar{z}^k, \bar{\xi}_k)\|^2 \mid \mathcal{F}_k\Big] \leq \alpha_k^2\gamma^2\mathbb{E}\Big[\|F\bar{z}^k\|^2 \mid \mathcal{F}_k\Big] + \alpha_k^2\gamma^2\sigma_F^2 \tag{C.8}$$

Combining (C.7), (C.8), and (C.5) yields

$$\begin{aligned}
&\mathbb{E}[\|z^{k+1} - z^\star\|^2 + A_{k+1}\|u^k\|^2 + B_{k+1}\|z^{k+1} - z^k\|^2 \mid \mathcal{F}_k] \\
&\quad\leq \|z^k - z^\star\|^2 + \Big(A_{k+1} + \alpha_k(\gamma^2 L_F^2(1+b) + \frac{1}{\varepsilon_k})\Big)\mathbb{E}[\|u^k\|^2 \mid \mathcal{F}_k] - \alpha_k\mu\|F(z^k)\|^2 \\
&\qquad+ \Big(\alpha_k(\gamma^2(\varepsilon_k - 1) - 2\gamma\rho) + \alpha_k^2\gamma^2\Big)\mathbb{E}\Big[\|F(\bar{z}^k)\|^2 \mid \mathcal{F}_k\Big] + \alpha_k^2\gamma^2\sigma_F^2 \\
&\qquad+ B_{k+1}\alpha_k^2\gamma^2\mathbb{E}\Big[\|F\bar{z}^k\|^2 \mid \mathcal{F}_k\Big] + B_{k+1}\alpha_k^2\gamma^2\sigma_F^2.
\end{aligned} \tag{C.9}$$

Further using (C.4) and denoting

$$\begin{aligned}
X_1^k &:= \alpha_k\Big(\gamma^2 L_F^2(1+b) + \tfrac{1}{\varepsilon_k}\Big) + A_{k+1}, \\
X_2^k &:= \alpha_k\Big(\gamma^2(\varepsilon_k - 1) - 2\rho\gamma + \alpha_k\,\gamma^2\Big)
\end{aligned}$$

leads to

$$\begin{aligned}
\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k \leq& -\alpha_k\mu\|F(z^k)\|^2 + \Big(X_1^k(1-\alpha_k)^2 - A_k\Big)\|u^{k-1}\|^2 \\
&+ \Big(2X_1^k(1-\alpha_k)^2\gamma^2 L_{\hat{F}}^2 - B_k\Big)\|z^k - z^{k-1}\|^2 + \Big(X_2^k + B_{k+1}\alpha_k^2\gamma^2\Big)\mathbb{E}\Big[\|F(\bar{z}^k)\|^2 \mid \mathcal{F}_k\Big] \\
&+ \Big(B_{k+1}\alpha_k^2 + \alpha_k^2 + 2X_1^k\alpha_k^2\Big)\gamma^2\sigma_F^2.
\end{aligned} \tag{C.10}$$

Having established (C.10), set $A_k = A$, $B_k = 2A\gamma^2 L_{\hat{F}}^2$, and $\varepsilon_k = \varepsilon$ to obtain by the law of total expectation that

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{k+1}] - \mathbb{E}[\mathcal{U}_k] \leq & -\alpha_k\mu\mathbb{E}\Big[\|F(z^k)\|^2\Big] + \Big(X_1^k(1-\alpha_k)^2 - A\Big)\mathbb{E}\Big[\|u^{k-1}\|^2\Big] \\ & + 2\gamma^2 L_{\hat{F}}^2\Big(X_1^k(1-\alpha_k)^2 - A\Big)\mathbb{E}\Big[\|z^k - z^{k-1}\|^2\Big] + \Big(X_2^k + 2A\gamma^4 L_{\hat{F}}^2\alpha_k^2\Big)\mathbb{E}\Big[\|F(\bar{z}^k)\|^2\Big] \\ & + \Big(2A\gamma^2 L_{\hat{F}}^2 + 1 + 2X_1^k\Big)\alpha_k^2\gamma^2\sigma_F^2. \end{aligned} \tag{C.11}$$

Let

$$A = \tfrac{1}{2}\Big((b+1)\gamma^2 L_F^2 + \tfrac{1}{\varepsilon}\Big).$$

Since $\alpha_k \in (0,1)$ it is of immediate verification that

$$X_1^k(1-\alpha_k)^2 - A \leq 0.$$

Note also that

$$\tfrac{1}{\alpha_k\gamma^2}\Big(X_2^k + 2A\gamma^4 L_{\hat{F}}^2\alpha_k^2\Big) = \varepsilon - 1 - 2\tfrac{\rho}{\gamma} + \alpha_k + 2A\gamma^2 L_{\hat{F}}^2\alpha_k \leq \varepsilon - 1 - \tfrac{2\rho}{\gamma} + \Big(1 + \Big((b+1)\gamma^2 L_F^2 + \tfrac{1}{\varepsilon}\Big)\gamma^2 L_{\hat{F}}^2\Big)\alpha_0$$

Setting the Young parameter $\varepsilon = \gamma L_{\hat{F}}\sqrt{\alpha_0}$ we obtain that $X_2^k + 2A\gamma^4 L_{\hat{F}}^2\alpha_k^2 \leq 0$ owing to (C.2).

On the other hand, the last term in (C.11) may be upper bounded by

$$\begin{aligned} 2A\gamma^2 L_{\hat{F}}^2 + 1 + 2X_1^k &= 1 + \Big((b+1)\gamma^2 L_F^2 + \tfrac{1}{\gamma L_{\hat{F}}\sqrt{\alpha_0}}\Big)\Big((\gamma^2 L_{\hat{F}}^2 + 1) + 2\alpha_k\Big) \\ &\leq 1 + \Big((b+1)\gamma^2 L_F^2 + \tfrac{1}{\gamma L_{\hat{F}}\sqrt{\alpha_0}}\Big)\Big((\gamma^2 L_{\hat{F}}^2 + 1) + 2\alpha_0\Big) = C. \end{aligned}$$

Thus, it follows from (C.11) that

$$\mathbb{E}[\mathcal{U}_{k+1}] - \mathbb{E}[\mathcal{U}_k] \leq -\alpha_k\mu\mathbb{E}\Big[\|F(z^k)\|^2\Big] + C\alpha_k^2\gamma^2\sigma_F^2.$$

Telescoping the above inequality completes the proof. □

Proof of Theorem 6.1. The theorem is obtained as a particular instantiation of Theorem C.2.

The condition in (C.1) can be rewritten as $b > \frac{\gamma^2 L_F^2}{1-\gamma^2 L_F^2}$. A reasonable choice is $b = \frac{2\gamma^2 L_F^2}{1-\gamma^2 L_F^2}$. Substituting back into μ we obtain

$$\mu = \gamma^2(1 - \gamma^2 L_F^2(1 + \tfrac{1-\gamma^2 L_F^2}{2\gamma^2 L_F^2})) = \tfrac{\gamma^2(1-\gamma^2 L_F^2)}{2} > 0. \tag{C.12}$$

Similarly, the choice of b is substituted into η and (C.2) of Theorem C.2.

The rate in (C.2) is further simplified by applying Lipschitz continuity of F from Assumption I*(i)* to $\|Fz^0\|^2 = \|Fz^0 - Fz^\star\|^2$. The proof is complete by observing that the guarantee on the weighted sum can be converted into an expectation over a sampled iterate in the style of Ghadimi & Lan (2013). □

Assumption VI (almost sure convergence)**.** *Let $d \in [0,1]$, $b > 0$. Suppose that the following holds*

(i) the diminishing sequence $(\alpha_k)_{k\in\mathbb{N}} \subset (0,1)$ satisfies the classical conditions

$$\textstyle\sum_{k=0}^\infty \alpha_k = \infty, \qquad \bar{\alpha} := \sum_{k=0}^\infty \alpha_k^2 < \infty;$$

(ii) letting $c_k := (1+b)\gamma^2 L_F^2 + \frac{1}{\gamma L_{\hat{F}}}\alpha_k^{-d}$ for all $k \geq 0$

$$\eta_k := \textstyle\sum_{\ell=k}^\infty \Big(c_l\alpha_l\Pi_{p=0}^\ell(1-\alpha_p)^2\Big) < \infty, \qquad \nu := \sum_{k=0}^\infty \eta_{k+1}\alpha_k^2\Big(\Pi_{p=0}^k \tfrac{1}{(1-\alpha_p)^2}\Big) < \infty, \tag{C.13}$$

and

$$\gamma L_{\hat{F}}\alpha_k^d + \alpha_k + 2\gamma^2 L_{\hat{F}}^2\alpha_k\eta_{k+1}\Pi_{p=0}^k\tfrac{1}{(1-\alpha_p)^2} \leq 1 + \tfrac{2\rho}{\gamma}. \tag{C.14}$$

Although at first look the above assumptions may appear involved, as shown in Theorem C.3 classical stepsize choice of $\frac{\alpha_0}{k+1}$ is sufficient to satisfy (C.13), and to ensure almost sure convergence provided that instead (C.16) holds. Note that with this choice as k goes to infinity, $\alpha_k \searrow 0$ and the deterministic range $\gamma + 2\rho > 0$ is obtained.

Theorem C.3 (almost sure convergence)**.** *Suppose that Assumptions I to III hold with $\Gamma = I$. Additionally, suppose the stepsize conditions in Assumptions V and VI. Then, the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 converges almost surely to some $z^\star \in \mathbf{zer}\, T$. Moreover, the following estimate holds*

$$\sum_{k=0}^{K} \frac{\alpha_k}{\sum_{j=0}^{K}\alpha_j}\mathbb{E}[\|F(z^k)\|^2] \le \frac{\|z^0 - z^\star\|^2 + \eta_0\gamma^2\|F(z^0)\|^2 + \bar{C}}{\mu\sum_{j=0}^{K}\alpha_j}, \tag{C.15}$$

where $\bar{C} = 2\gamma^2\sigma_F^2((\gamma^2 L_{\hat{F}}^2 + 1)\nu + \bar{\alpha}(\frac{1}{2} + (b+1)\gamma^2 L_F^2 + \frac{1}{\gamma L_{\hat{F}}}))$ is finite.

In particular, if $\alpha_k = \frac{1}{k+r}$ for any positive natural number r, and $d = 1$, then Assumption VI(ii) can be replaced by

$$(\gamma L_{\hat{F}} + 1)\alpha_k + \Big((1+b)\gamma^4 L_F^4 \alpha_{k+1} + \gamma L_{\hat{F}}\Big)(\alpha_{k+1} + 1)\alpha_{k+1} \le 1 + \tfrac{2\rho}{\gamma}. \tag{C.16}$$

Proof of Theorem C.3 *(almost sure convergence).* Having established (C.10), let $B_k = 2A_k\gamma^2 L_{\hat{F}}^2$ such that

$$\Big(2X_1^k(1-\alpha_k)^2\gamma^2 L_{\hat{F}}^2 - B_k\Big)\|z^k - z^{k-1}\|^2 = 2\gamma^2 L_{\hat{F}}^2\Big(X_1^k(1-\alpha_k)^2 - A_k\Big)\|z^k - z^{k-1}\|^2. \tag{C.17}$$

In what follows we show that it is sufficient to ensure

$$X_1^k(1-\alpha_k)^2 \le A_k, \quad X_2^k + 2A_{k+1}\gamma^4 L_{\hat{F}}^2\alpha_k^2 \le 0, \tag{C.18}$$

resulting in the inequality

$$\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k \le -\alpha_k\mu\|F(z^k)\|^2 + \Big(2A_{k+1}\gamma^2 L_{\hat{F}}^2 + 1 + 2X_1^k\Big)\alpha_k^2\gamma^2\sigma_F^2. \tag{C.19}$$

A reasonable choice for the Young parameter ε_k is to choose

$$\varepsilon_k = \gamma L_{\hat{F}}\alpha_k^d \quad \text{for some} \quad d \in [0,1]. \tag{C.20}$$

The rational for this choice will become more clear in what follows.

The first inequality in (C.18) is linear and we can solve it to equality by Lemma C.1. Letting c_k and η_k be as in Assumption VI*(ii)*

$$A_{k+1} = \left(\Pi_{p=0}^{k}\frac{1}{(1-\alpha_p)^2}\right)\left(A_0 - \sum_{\ell=0}^{k}\Big(c_l\alpha_l\Pi_{p=0}^{\ell}(1-\alpha_p)^2\Big)\right) = \eta_{k+1}\Pi_{p=0}^{k}\frac{1}{(1-\alpha_p)^2} \tag{C.21}$$

which would ensure $A_k \ge 0$ for all k. Let

$$A_0 := \sum_{\ell=0}^{\infty}\Big(c_l\alpha_l\Pi_{p=0}^{\ell}(1-\alpha_p)^2\Big) = \eta_0 \overset{(C.13)}{<} \infty, \quad \text{and} \quad \nu = \sum_{k=0}^{\infty} A_{k+1}\alpha_k^2 \overset{(C.13)}{<} \infty. \tag{C.22}$$

Therefore, assumptions (C.13) and (C.14) (which is a restatement of the conditions in (C.18)) are sufficient for ensuring (C.19). Substituting X_1^k and A_{k+1} in (C.19) yields

$$\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k \le -\alpha_k\mu\|F(z^k)\|^2 + \xi_k, \tag{C.23}$$

where $\xi_k = 2\Big(A_{k+1}(\gamma^2 L_{\hat{F}}^2 + 1) + \frac{1}{2} + (b+1)\gamma^2 L_F^2\alpha_k + \frac{1}{\gamma L_{\hat{F}}}\alpha_k^{1-d}\Big)\alpha_k^2\gamma^2\sigma_F^2$. By Assumption VI we have that

$$\begin{aligned}\textstyle\sum_{k=0}^{\infty}\xi_k &= 2\gamma^2\sigma_F^2\left((\gamma^2 L_{\hat{F}}^2 + 1)\sum_{k=0}^{\infty}A_{k+1}\alpha_k^2 + \sum_{k=0}^{\infty}\tfrac{\alpha_k^2}{2} + (b+1)\gamma^2 L_F^2\sum_{k=0}^{\infty}\alpha_k^3 + \tfrac{1}{\gamma L_{\hat{F}}}\sum_{k=0}^{\infty}\alpha_k^{3-d}\right)\\ &\le 2\gamma^2\sigma_F^2\left((\gamma^2 L_{\hat{F}}^2 + 1)\sum_{k=0}^{\infty}A_{k+1}\alpha_k^2 + \Big(\tfrac{1}{2} + (b+1)\gamma^2 L_F^2 + \tfrac{1}{\gamma L_{\hat{F}}}\Big)\sum_{k=0}^{\infty}\alpha_k^2\right) < \infty\end{aligned}$$

where we used the fact that $d \le 1$, (C.22), and Assumption VI*(i)*. The claimed convergence result follows by the Robbins-Siegmund supermartingale theorem (Bertsekas, 2011, Prop. 2) and standard arguments as in (Bertsekas, 2011, Prop. 9).

The claimed rate follows by taking total expectation and summing the above inequality over k and noting that initial iterates were set as $\bar{z}^{-1} = z^{-1} = z^0$.

To provide an instance of the sequence $(\alpha_k)_{k\in\mathbb{N}}$ that satisfy the assumptions, let r denote a positive natural number and set

$$\alpha_k = \tfrac{1}{k+r}. \tag{C.24}$$

Then,

$$\Pi_{p=0}^{\ell}(1-\alpha_p)^2 = \Pi_{p=0}^{\ell}(\tfrac{p+r-1}{p+r})^2 = \tfrac{(r-1)^2}{(\ell+r)^2} = (r-1)^2\alpha_\ell^2,$$

and for any $K \geq 0$

$$\sum_{\ell=0}^{K}\Big(c_\ell\alpha_\ell\Pi_{p=0}^{\ell}(1-\alpha_p)^2\Big) = \sum_{\ell=0}^{K}\tfrac{(r-1)^2}{(\ell+r)^3}c_\ell.$$

Plugging the value of c_ℓ and ε_k from (C.21) we obtain that A_0 is finite valued since $\sum_{\ell=0}^{\infty}\frac{1}{(\ell+r)^3\varepsilon_\ell} = \sum_{\ell=0}^{\infty}\frac{1}{(\ell+r)^{3-d}} < \infty$ owing to the fact that $d \leq 1$.

Moreover,

$$A_{k+1} = \frac{(k+r)^2}{(r-1)^2}\Bigg(A_0 - \sum_{\ell=0}^{k}\Big(\tfrac{(r-1)^2}{(\ell+r)^3}c_\ell\Big)\Bigg) = (k+r)^2\sum_{\ell=k+1}^{\infty}\tfrac{1}{(\ell+r)^3}c_\ell = \tfrac{1}{\alpha_k^2}\sum_{\ell=k+1}^{\infty}\alpha_\ell^3 c_\ell \tag{C.25}$$

On the other hand, for $e > 1$ we have the following bound

$$\sum_{\ell=k+1}^{\infty}\alpha_\ell^e \leq \tfrac{1}{(k+1+r)^e} + \int_{k+1}^{\infty}\tfrac{1}{(x+r)^e}dx = \tfrac{1}{(k+1+r)^e} + \tfrac{1}{(e-1)(k+1+r)^{e-1}}. \tag{C.26}$$

Therefore, it follows from (C.25) that

$$\begin{aligned} A_{k+1}\alpha_k &= \tfrac{1}{\alpha_k}\sum_{\ell=k+1}^{\infty}(\alpha_\ell^3(1+b)\gamma^2L_F^2 + \tfrac{1}{\gamma L_F}\alpha_\ell^{3-d}) \\ \text{(C.26)} &\leq \Big((1+b)\gamma^2L_F^2\tfrac{1}{2(k+1+r)} + \tfrac{1}{\gamma L_{\hat{F}}}\tfrac{1}{(2-d)(k+1+r)^{1-d}}\Big)\Big(\tfrac{1}{k+1+r}+1\Big)\tfrac{1}{k+1+r} \\ &= \Big(\tfrac{1+b}{2}\gamma^2L_F^2\alpha_{k+1} + \tfrac{1}{\gamma L_{\hat{F}}(2-d)}\alpha_{k+1}^{1-d}\Big)(\alpha_{k+1}+1)\alpha_{k+1}. \end{aligned} \tag{C.27}$$

In turn, this inequality ensures that ν as defined in Assumption VI*(ii)* is finite. To see this note that

$$\nu = \textstyle\sum_{k=0}^{\infty}A_{k+1}\alpha_k^2 \overset{\text{(C.27)}}{\leq} \sum_{k=0}^{\infty}\Big(\frac{1+b}{2}\gamma^2L_F^2\alpha_{k+1} + \frac{1}{\gamma L_{\hat{F}}(2-d)}\alpha_{k+1}^{1-d}\Big)(\alpha_{k+1}+1)\alpha_{k+1}\alpha_k \leq \delta\sum_{k=0}^{\infty}\alpha_k^2 < \infty,$$

where in the last two inequalities Assumption VI*(i)* was used.

It remains to confirm the second inequality in (C.18). With the choice of α_k and ε_k as in (C.24) and (C.20) we have

$$\begin{aligned} \tfrac{1}{\alpha_k\gamma^2}(X_2 + 2A_{k+1}\gamma^4L_{\hat{F}}^2\alpha_k^2) &= \gamma L_{\hat{F}}\alpha_k^d - 1 - \tfrac{2\rho}{\gamma} + \alpha_k + 2A_{k+1}\gamma^2L_{\hat{F}}^2\alpha_k \\ \text{(C.27)} &\leq \gamma L_{\hat{F}}\alpha_k^d + \alpha_k + 2\gamma^2L_{\hat{F}}^2\Big(\tfrac{1+b}{2}\gamma^2L_F^2\alpha_{k+1} + \tfrac{1}{\gamma L_{\hat{F}}(2-d)}\alpha_{k+1}^{1-d}\Big)(\alpha_{k+1}+1)\alpha_{k+1} - 1 - \tfrac{2\rho}{\gamma}. \end{aligned}$$

It follows that with $d = 1$ the assumption (C.16) is sufficient to ensure that the second condition in (C.18) holds. □

Proof of Theorem 6.3 *(almost sure convergence).* The result is a restatement of the special case in Theorem C.3 where $\alpha_k = \frac{1}{k+r}$. We proceed similarly to the proof of Theorem 6.1.

The condition in (C.1) can be rewritten as $b > \frac{\gamma^2L_F^2}{1-\gamma^2L_F^2}$. A reasonable choice is $b = \frac{2\gamma^2L_F^2}{1-\gamma^2L_F^2}$. The choice of b is substituted into (C.1), (C.16) and $\bar{C}$ of Theorem C.3. This completes the proof.

□

D Proof for constrained case

Theorem D.1. *Suppose that Assumptions I to III hold with $\Gamma = I$. Moreover, suppose that $\alpha_k \in (0,1)$ and for positive parameters ε and b the following holds,*

$$\mu := 1+(1+b)\tfrac{2\rho}{\gamma} - \frac{1}{\varepsilon(1-\gamma^2L_F^2)} - \alpha_0(1+b)(1+2A\gamma^2L_{\hat{F}}^2) > 0 \quad \textit{and} \quad 1-2(1+1/b)\alpha_0\gamma^2L_{\hat{F}}^2 > 0 \tag{D.1}$$

where $A \geq \frac{\varepsilon+(1+1/b)(\alpha_0-\frac{2\rho}{\gamma})}{1-2(1+1/b)\alpha_0\gamma^2 L_{\hat{F}}^2}$. *Consider the sequence* $(z^k)_{k\in\mathbb{N}}$ *generated by Algorithm 2. Then, the following estimate holds for all* $z^\star \in \mathcal{S}^\star$

$$\sum_{k=0}^{K} \frac{\alpha_k}{\sum_{j=0}^{K}\alpha_j}\mathbb{E}[\|H\bar{z}^k - Hz^k\|^2] \leq \frac{\mathbb{E}[\|z^0 - z^\star\|^2] + A\mathbb{E}[\|h^{-1} - Hz^{-1}\|^2] + C\gamma^2\sigma_F^2\sum_{j=0}^{K}\alpha_j^2}{\mu\sum_{j=0}^{K}\alpha_j}$$

where $C = 1 + 2A(1+\gamma^2 L_{\hat{F}}^2) + 2\alpha_0(\varepsilon + (1+1/b)(\alpha_0 - \frac{2\rho}{\gamma} + 2\alpha_0\gamma^2 L_{\hat{F}}^2 A))$.

Proof of Theorem D.1. We rely on the following potential function,

$$\mathcal{U}_{k+1} := \|z^{k+1} - z^\star\|^2 + A_{k+1}\|h^k - Hz^k\|^2 + B_{k+1}\|z^{k+1} - z^k\|^2,$$

where $(A_k)_{k\in\mathbb{N}}$ and $(B_k)_{k\in\mathbb{N}}$ are positive scalar parameters to be identified.

We will denote $\hat{\bar{H}}_k := \bar{z}^k - \gamma\hat{F}(\bar{z}^k, \bar{\xi}_k)$, so that $z^{k+1} = z^k - \alpha_k(h^k - \hat{\bar{H}}_k)$. Then, expanding one step,

$$\|z^{k+1} - z^\star\|^2 = \|z^k - z^\star\|^2 - 2\alpha_k\langle h^k - \hat{\bar{H}}_k, z^k - z^\star\rangle + \alpha_k^2\|h^k - \hat{\bar{H}}_k\|^2. \tag{D.2}$$

Recall that $Hz := z - \gamma Fz$ in the deterministic case. In the Algorithm 2, h^k estimates Hz^k. Let us quantify how good this estimation is.

$$\begin{aligned}
h^k - Hz^k &= \gamma Fz^k - \gamma\hat{F}(z^k,\xi_k) + (1-\alpha_{k-1})(h^{k-1} - z^{k-1} + \gamma\hat{F}(z^{k-1},\xi_k)) \\
\|h^k - Hz^k\|^2 &= (1-\alpha_{k-1})^2\|h^{k-1} - z^{k-1} + \gamma Fz^{k-1}\|^2 \\
&\quad + \|\gamma Fz^k - \gamma\hat{F}(z^k,\xi_k) + (1-\alpha_{k-1})(\gamma\hat{F}(z^{k-1},\xi_k) - \gamma Fz^{k-1})\|^2 \\
&\quad + 2(1-\alpha_{k-1})\langle h^{k-1} - z^{k-1} + \gamma Fz^{k-1}, \\
&\qquad\qquad \gamma Fz^k - \gamma\hat{F}(z^k,\xi_k) + (1-\alpha_{k-1})(\gamma\hat{F}(z^{k-1},\xi_k) - \gamma Fz^{k-1})\rangle
\end{aligned}$$

In the scalar product, the left term is known when z^k is known and the right term has an expectation equal to 0 by Assumption II*(i)* when z^k is known. Thus using the fact that the second moment is larger than the variance, we can go on as

$$\begin{aligned}
\mathbb{E}[\|h^k - Hz^k\|^2 \mid \mathcal{F}_k] &\leq (1-\alpha_{k-1})^2\|h^{k-1} - H_{k-1}(z^{k-1})\|^2 \\
&\quad + \mathbb{E}[2(1-\alpha_{k-1})^2\gamma^2\|\hat{F}(z^k,\xi_k) - \hat{F}(z^{k-1},\xi_k)\|^2 \mid \mathcal{F}_k] \\
&\quad + \mathbb{E}[2\alpha_{k-1}^2\gamma^2\|Fz^k - \hat{F}(z^k,\xi_k)\|^2 \mid \mathcal{F}_k] \\
&\leq (1-\alpha_{k-1})^2\|h^{k-1} - H_{k-1}(z^{k-1})\|^2 + 2(1-\alpha_{k-1})^2 L_{\hat{F}}^2\gamma^2\|z^k - z^{k-1}\|^2 + 2\alpha_{k-1}^2\gamma^2\sigma_F^2
\end{aligned} \tag{D.3}$$

where we have used Assumption II*(ii)* and Assumption III.

We continue with the inner term in (D.2).

$$\begin{aligned}
-\mathbb{E}[\langle h^k - \hat{\bar{H}}_k, z^k - z^\star\rangle \mid \mathcal{F}_k] &= \langle -h^k + H\bar{z}^k, z^k - z^\star\rangle \\
&= \langle -h^k + H\bar{z}^k, z^k - \bar{z}^k\rangle + \langle -h^k + H\bar{z}^k, \bar{z}^k - z^\star\rangle \\
&= \langle -h^k + Hz^k, z^k - \bar{z}^k\rangle - \langle Hz^k - H\bar{z}^k, z^k - \bar{z}^k\rangle + \langle -h^k + H\bar{z}^k, \bar{z}^k - z^\star\rangle
\end{aligned} \tag{D.4}$$

By definition of $\bar{z}^k$ in Step 2.3, we have $h^k \in \bar{z}^k + \gamma A(\bar{z}^k)$, so that $\frac{1}{\gamma}(h^k - H\bar{z}^k) \in F(\bar{z}^k) + A(\bar{z}^k)$. Hence, using the weak MVI from Assumption I*(iii)*,

$$\langle h^k - H\bar{z}^k, \bar{z}^k - z^\star\rangle \geq \frac{\rho}{\gamma}\|h^k - H\bar{z}^k\|^2 . \tag{D.5}$$

To bound the second term of (D.4), we use that H is $\frac{1}{2}$-cocoercive, i.e. for all $z, z' \in \mathbb{R}^n$

$$\begin{aligned}
\langle Hz - Hz', z - z'\rangle &= \langle Hz - Hz', Hz - Hz' + \gamma Fz' - \gamma Fz\rangle \\
&= \tfrac{1}{2}\|Hz - Hz'\|^2 - \tfrac{\gamma^2}{2}\|Fz' - Fz\|^2 + \tfrac{1}{2}\|z' - z\|^2 \\
&\geq \tfrac{1}{2}\|Hz - Hz'\|^2,
\end{aligned} \tag{D.6}$$

under Assumption I*(i)* when $\gamma \leq 1/L_F$. Using (D.5) and (D.6) in (D.4) leads to the following inequality, true for any $\varepsilon_k > 0$:

$$-\mathbb{E}[\langle h^k - \hat{\bar{H}}_k, z^k - z^\star\rangle \mid \mathcal{F}_k] \leq \frac{\varepsilon_k}{2}\|h^k - Hz^k\|^2 + \frac{1}{2\varepsilon_k}\|\bar{z}^k - z^k\|^2 - \frac{1}{2}\|Hz^k - H\bar{z}^k\|^2 - \frac{\rho}{\gamma}\|h^k - H\bar{z}^k\|^2 .$$

To majorize the term $\|\bar{z}^k - z^k\|^2$, we may use the fact that

$$\|Hz - Hz'\|^2 = \|z - z' - \gamma(Fz + Fz')\|^2 \geq \|z - z'\|^2 - \gamma^2\|Fz - Fz'\|^2 \geq (1 - \gamma^2 L_F^2)\|z - z'\|^2 \ .$$

Hence,

$$-\mathbb{E}[\langle h^k - \hat{H}_k, z^k - z^\star\rangle \mid \mathcal{F}_k] \leq \tfrac{\varepsilon_k}{2}\|h^k - Hz^k\|^2 + \Big(\tfrac{1}{2\varepsilon_k(1-\gamma^2 L_{\hat{F}}^2)} - \tfrac{1}{2}\Big)\|Hz^k - H\bar{z}^k\|^2 - \tfrac{\rho}{\gamma}\|h^k - H\bar{z}^k\|^2 \ . \tag{D.7}$$

The third term in (D.2) is bounded by

$$\alpha_k^2\mathbb{E}[\|h^k - \hat{H}_k\|^2 \mid \mathcal{F}_k] = \alpha_k^2\|h^k - H\bar{z}^k\|^2 + \alpha_k^2\gamma^2\mathbb{E}[\|F\bar{z}^k - \hat{F}(\bar{z}^k, \bar{\xi}_k)\|^2 \mid \mathcal{F}_k] \leq \alpha_k^2\|h^k - H\bar{z}^k\|^2 + \alpha_k^2\gamma^2\sigma_F^2 \tag{D.8}$$

Combined with the update rule, (D.8) can also be used to bound the difference of iterates

$$\mathbb{E}[\|z^{k+1} - z^k\|^2 \mid \mathcal{F}_k] = \mathbb{E}[\alpha_k^2\|h^k - \hat{H}_k\|^2 \mid \mathcal{F}_k] \leq \alpha_k^2\|h^k - H_k(\bar{z}^k)\|^2 + \alpha_k^2\gamma^2\sigma_F^2 \tag{D.9}$$

Using (D.2), (D.7), (D.8) and (D.9) we have,

$$\begin{aligned}
\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] &\leq \|z^k - z^\star\|^2 + (A_{k+1} + \alpha_k\varepsilon_k)\|h^k - Hz^k\|^2 + \alpha_k\Big(\tfrac{1}{\varepsilon_k(1-\gamma^2 L_{\hat{F}}^2)} - 1\Big)\|Hz^k - H\bar{z}^k\|^2 \\
&\quad + \alpha_k(\alpha_k - \tfrac{2\rho}{\gamma} + \alpha_k B_{k+1})\|h^k - H\bar{z}^k\|^2 + \alpha_k^2(1 + B_{k+1})\gamma^2\sigma_F^2 \\
&\leq \|z^k - z^\star\|^2 + (A_{k+1} + \alpha_k(\varepsilon_k + (1 + 1/b)(\alpha_k - \tfrac{2\rho}{\gamma} + \alpha_k B_{k+1})))\|h^k - Hz^k\|^2 \\
&\quad + \alpha_k\Big(\tfrac{1}{\varepsilon_k(1-\gamma^2 L_{\hat{F}}^2)} - 1 + (1 + b)(\alpha_k - \tfrac{2\rho}{\gamma} + \alpha_k B_{k+1})\Big)\|Hz^k - H\bar{z}^k\|^2 \\
&\quad + \alpha_k^2(1 + B_{k+1})\gamma^2\sigma_F^2,
\end{aligned} \tag{D.10}$$

where the last inequality follows from Young's inequality with positive b. By defining

$$\begin{aligned}
X_k^1 &\coloneqq A_{k+1} + \alpha_k(\varepsilon_k + (1 + 1/b)(\alpha_k - \tfrac{2\rho}{\gamma} + \alpha_k B_{k+1})) \\
X_k^2 &\coloneqq \alpha_k\Big(\tfrac{1}{\varepsilon_k(1-\gamma^2 L_{\hat{F}}^2)} - 1 + (1 + b)(\alpha_k - \tfrac{2\rho}{\gamma} + \alpha_k B_{k+1})\Big)
\end{aligned} \tag{D.11}$$

and applying (D.3), we finally obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k &\leq X_k^2\|Hz^k - H\bar{z}^k\|^2 \\
&\quad + (X_k^1(1 - \alpha_{k-1})^2 - A_k)\|h^{k-1} - H_{k-1}(z^{k-1})\|^2 \\
&\quad + (2X_k^1(1 - \alpha_{k-1})^2\gamma^2 L_{\hat{F}}^2 - B_k)\|z^k - z^{k-1}\|^2 \\
&\quad + 2X_k^1\alpha_{k-1}^2\gamma^2\sigma_F^2 + \alpha_k^2(1 + B_{k+1})\gamma^2\sigma_F^2,
\end{aligned} \tag{D.12}$$

We can pick $B_k = 2\gamma^2 L_{\hat{F}}^2 A_k$ in which case, to get a recursion, we only require the following.

$$X_k^1(1 - \alpha_{k-1})^2 \leq A_k \quad \text{and} \quad X_k^2 < 0 \tag{D.13}$$

Set $A_k = A$, $\varepsilon_k = \varepsilon$. For the first inequality of (D.13), since $(1 - \alpha_k)^2 \leq (1 - \alpha_k)$, the terms involving A are bounded as

$$\begin{aligned}
(1 - \alpha_k)^2(1 + 2\gamma^2 L_{\hat{F}}^2(1 + 1/b)\alpha_k^2)A - A &\leq (1 - \alpha_k)(1 + 2\gamma^2 L_{\hat{F}}^2(1 + 1/b)\alpha_k^2)A - A \\
&\leq (1 - \alpha_k)(2\gamma^2 L_{\hat{F}}^2(1 + 1/b)\alpha_k^2)A - \alpha_k A \\
&\leq -\alpha_k(1 - 2\gamma^2 L_{\hat{F}}^2(1 + 1/b)\alpha_0)A
\end{aligned} \tag{D.14}$$

where the last inequality follows from $(1 - \alpha_k) \leq 1$ and $\alpha_k \leq \alpha_0$. Thus to satisfy the first inequality of (D.13) it suffice to pick

$$A \geq \frac{\varepsilon + (1 + 1/b)(\alpha_0 - \frac{2\rho}{\gamma})}{1 - 2(1 + 1/b)\alpha_0\gamma^2 L_{\hat{F}}^2} \tag{D.15}$$

where $1 - 2(1 + 1/b)\alpha_0\gamma^2 L_{\hat{F}}^2 > 0$ as also stated in (D.1) together with the second equality of (D.13).

The noise term in (D.12) can be made independent of k by using $\alpha_k \leq \alpha_0$ as follows

$$\begin{aligned}
2X_k^1 + 1 + B_{k+1} &= 1 + 2A(1 + \gamma^2 L_{\hat{F}}^2) + 2\alpha_k(\varepsilon + (1 + 1/b)(\alpha_k - \tfrac{2\rho}{\gamma} + 2\alpha_k\gamma^2 L_{\hat{F}}^2 A)) \\
&\leq 1 + 2A(1 + \gamma^2 L_{\hat{F}}^2) + 2\alpha_0(\varepsilon + (1 + 1/b)(\alpha_0 - \tfrac{2\rho}{\gamma} + 2\alpha_0\gamma^2 L_{\hat{F}}^2 A)) = C.
\end{aligned} \tag{D.16}$$

Thus it follows from (D.12) that

$$\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k \leq \alpha_k\Big(\frac{1}{\varepsilon(1-\gamma^2 L_F^2)} - 1 + (1+b)(\alpha_k - \tfrac{2\rho}{\gamma} + 2\alpha_k\gamma^2 L_{\hat{F}}^2 A)\Big)\|Hz^k - H\bar{z}^k\|^2 + \alpha_k^2 C\gamma^2\sigma_F^2. \tag{D.17}$$

The result is obtained by total expectation and summing the above inequality while noting that the initial iterate were set as $z^{-1} = z^0$. □

Proof of Theorem 7.1. The theorem is a specialization of Theorem D.1 with $\Gamma = \gamma I$ and a particular a choice of b and ε. The requirement (D.1) can be rewritten as,

$$b > \frac{2\alpha_0\gamma^2 L_{\hat{F}}^2}{1 - 2\alpha_0\gamma^2 L_{\hat{F}}^2}, \tag{D.18}$$

which is satisfied by $b = \frac{4\alpha_0\gamma^2 L_{\hat{F}}^2}{1-2\alpha_0\gamma^2 L_{\hat{F}}^2}$. We substitute in the choice of b, $\varepsilon = \frac{1}{(1-\gamma^2 L_F^2)\sqrt{\alpha_0}}$ and denotes $\eta := A$.

The proof is complete by converting the weighted sum into an expectation over a sampled iterate in the style of Ghadimi & Lan (2013). □

E Proof for nonlinearly preconditioned PDHG through asymmetric preconditioner

Theorem E.1. *Suppose that Assumption I(ii) to II and IV hold. Moreover, suppose that $\alpha_k \in (0,1)$, $\theta \in [0,1]$ and the following holds,*

$$\begin{aligned} \mu &:= 1 + (1+b)\tfrac{2\rho}{\bar{\gamma}} - \frac{1}{\varepsilon(1-L_M^2)} - \alpha_0(1+b)(1 + 2(\hat{c}_1 + \hat{c}_4 A) > 0 \quad \textit{and} \\ \nu &:= 1 - 4\hat{c}_2\alpha_0 - (1 + 4\hat{c}_2\alpha_0^2)(2\alpha_0(1 + 1/b)(\hat{c}_1 + \hat{c}_4)) > 0 \end{aligned} \tag{E.1}$$

where $\bar{\gamma}$ denotes the smallest eigenvalue of Γ, $A \geq \frac{\varepsilon + (1+1/b)(\alpha_0 - \frac{2\rho}{\bar{\gamma}})}{\nu}$ and

$$\hat{c}_1 := \|\Gamma(D_{\widehat{xz}} + D_{\widehat{yz}}) + \Gamma_2 D_{\widehat{yy}}\|, \quad \hat{c}_2 := \|\Gamma_1 D_{\widehat{yx}}\|, \quad \hat{c}_3 := \|D_{\widehat{xz}}\Gamma\|, \quad \hat{c}_4 = 2\hat{c}_2(1+\hat{c}_3)$$
$$L_M^2 := \mathbf{max}\{\|D_{xx}\Gamma_1\| + \|D_{xy}\Gamma_2\|, \|D_{yx}\Gamma_1\| + \|D_{yy}\Gamma_2\|\}.$$

Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 3. Then, the following estimate holds for all $z^\star \in \mathcal{S}^\star$

$$\sum_{k=0}^{K} \tfrac{\alpha_k}{\sum_{j=0}^K \alpha_j}\mathbb{E}[\|H_{z^k}(\bar{z}^k) - H_{z^k}(z^k)\|_\Gamma^2] \leq \frac{\mathbb{E}[\|z^0 - z^\star\|_{\Gamma^{-1}}^2] + A\mathbb{E}[\|\hat{z}^{-1} - S_{z^{-1}}(z^{-1};\bar{z}^{-1})\|_\Gamma^2] + C\sigma_F^2\sum_{j=0}^K \alpha_j^2}{\mu\sum_{j=0}^K \alpha_j}$$

where $C := 2(A + \alpha_0(\varepsilon + (1+b^{-1})(\alpha_0 - 2\frac{\rho}{\bar{\gamma}} + \alpha_0 2(\hat{c}_1 + \hat{c}_4 A)))(1 + 2\hat{c}_2) + 1 + 2(\hat{c}_1 + \hat{c}_4)A$.

Lemma E.2. *Suppose Assumption II for $F(z) = (\nabla_x\varphi(z), -\nabla_y\varphi(z))$ as in (8.6). Let $\hat{M}_u$ and M_u be as defined in (E.4) and $\hat{Q}_u$ and Q_u as in (8.8) with $\theta \in [0,1]$. Then the following holds*

$$\begin{aligned} &\mathbb{E}_\xi[\hat{M}_z(z,\xi)] = M_z(z), \quad \mathbb{E}_\xi[\hat{Q}_z(z',\xi)] = Q_z(z') \quad \textit{and} \\ &\mathbb{E}_{\xi,\xi'}[\|M_z(z) - \hat{M}_z(z,\xi) + Q_z(z') - \hat{Q}_z(z',\xi')\|_\Gamma^2] \leq \sigma_F^2. \end{aligned} \tag{E.2}$$

Proof. For all $z, z' \in \mathbb{R}^n$ where $z = (x, y)$ and $z' = (x', y')$

$$\begin{aligned}
&\mathbb{E}_{\xi,\xi'}[\|M_z(z) - \hat{M}_z(z,\xi) + Q_z(z') - \hat{Q}_z(z',\xi')\|_\Gamma^2] \\
&= \mathbb{E}_{\xi,\xi'}\left[\left\|\begin{pmatrix} \nabla_x\hat{\varphi}(z,\xi) - \nabla_x\varphi(z) \\ (1-\theta)(\nabla_y\hat{\varphi}(z,\xi) - \nabla_y\varphi(z')) + \theta(\nabla_y\hat{\varphi}(x',y,\xi') - \nabla_y\varphi(x',y)) \end{pmatrix}\right\|_\Gamma^2\right] \\
&= \mathbb{E}_{\xi,\xi'}\left[\left\|\begin{pmatrix} (1-\theta)(\nabla_x\hat{\varphi}(z,\xi) - \nabla_x\varphi(z)) + \theta(\nabla_x\hat{\varphi}(z,\xi) - \nabla_x\varphi(z)) \\ (1-\theta)(\nabla_y\hat{\varphi}(z,\xi) - \nabla_y\varphi(z')) + \theta(\nabla_y\hat{\varphi}(\bar{x},y,\xi') - \nabla_y\varphi(\bar{x},y)) \end{pmatrix}\right\|_\Gamma^2\right] \\
\text{(Assumption II}\textit{(i)}\text{)} &\le (1-\theta)^2\mathbb{E}_{\xi,\xi'}\left[\left\|\begin{pmatrix} \nabla_x\hat{\varphi}(z,\xi) - \nabla_x\varphi(z) \\ \nabla_y\hat{\varphi}(z,\xi) - \nabla_y\varphi(z') \end{pmatrix}\right\|_\Gamma^2\right] + \theta^2\mathbb{E}_{\xi,\xi'}\left[\left\|\begin{pmatrix} \nabla_x\hat{\varphi}(z,\xi) - \nabla_x\varphi(z) \\ \nabla_y\hat{\varphi}(\bar{x},y,\xi') - \nabla_y\varphi(\bar{x},y) \end{pmatrix}\right\|_\Gamma^2\right] \\
\text{(Assumption II}\textit{(ii)}\text{)} &\le (1-\theta)^2\sigma_F^2 + \theta^2\sigma_F^2 \le \sigma_F^2.
\end{aligned} \tag{E.3}$$

Unbiasedness follows immediately through Assumption II*(i)*. This completes the proof. □

Proof of Theorem E.1. The proof relies on tracking the two following important quantities instead of F and $\hat{F}$

$$M_u(z) := F(z) - Q_u(z) \quad \text{and} \quad \hat{M}_u(z,\xi) := \hat{F}(z,\xi) - \hat{Q}_u(z,\xi). \tag{E.4}$$

We will compare the stochastic algorithm with its deterministic counterpart, so we introduce

$$H_{z^k}(z) := P_{z^k}(z) - F(z) \tag{E.5a}$$

$$\bar{G}(z) := (P_{z^k} + A)^{-1}(H_{z^k}(z)) \tag{E.5b}$$

$$G(z) := z - \alpha_k\Gamma\Big(H_{z^k}(z) - H_{z^k}(\bar{G}(z))\Big). \tag{E.5c}$$

The proof first derive results for the deterministic operator G and then shows that z^{k+1} behaves similarly to $G(z^k)$ when α_k is small enough, even if Γ, which also appears inside the preconditioner $\hat{P}_u(\cdot,\xi)$, remains large.

In what follows we first establish that the mapping H_{z^k} is $1/2$-cocoercive along the points $z^k, \bar{G}z^k$ generated by the algorithm. By expanding

$$H_{z^k}(z^k) - H_{z^k}(\bar{G}z^k) = \Gamma^{-1}(z^k - \bar{G}z^k) - (M_{z^k}(z^k) - M_{z^k}(\bar{G}z^k)). \tag{E.6}$$

where $M_{z^k}(z) := F(z) - Q_{z^k}(z)$. Using this we have

$$\begin{aligned}
\langle H_{z^k}(\bar{G}z^k) - H_{z^k}(z^k), \bar{G}z^k - z^k\rangle &= \langle H_{z^k}(\bar{G}z^k) - H_{z^k}(z^k), H_{z^k}(\bar{G}z^k) - H_{z^k}(z^k) - (M_{z^k}(z^k) - M_{z^k}(\bar{G}z^k))\rangle_\Gamma \\
\text{(E.6)} &= \tfrac{1}{2}\|H_{z^k}(\bar{G}z^k) - H_{z^k}(z^k)\|_\Gamma^2 + \tfrac{1}{2}\|\bar{G}z^k - z^k\|_{\Gamma^{-1}}^2 - \tfrac{1}{2}\|M_{z^k}(z^k) - M_{z^k}(\bar{G}z^k)\|_\Gamma^2 \\
\text{Requirements IV.A2 and IV.A3} &\ge \tfrac{1}{2}\|H_{z^k}(\bar{G}z^k) - H_{z^k}(z^k)\|_\Gamma^2
\end{aligned} \tag{E.7}$$

We now go back to our stochastic template algorithm defined in (8.4). We will denote $\hat{\hat{H}}_k := \hat{P}_k(\bar{z}^k, \bar{\xi}_k) - \hat{F}(\bar{z}^k, \bar{\xi}_k)$, so that $z^{k+1} = z^k - \alpha_k\Gamma(h^k - \hat{\hat{H}}_k)$. We will further need the following change of variables to later be able to apply weak MVI (see Appendix E.1):

$$\begin{aligned}
s^k &= h^k - \hat{Q}_{z^k}(\bar{z}^k, \xi_k') \\
\hat{\hat{S}}_k &= \hat{\hat{H}}_k - \hat{Q}_{z^k}(\bar{z}^k, \xi_k') \\
S_u(\bar{z}) &= H_u(\bar{z}) - Q_u(\bar{z}) \\
S_u(z;\bar{z}) &= H_u(z) - Q_u(\bar{z})
\end{aligned} \tag{E.8}$$

where $Q_u(z)$ and H_u are as defined in Section 8.

In contrast with the unconstrained smooth case we will rely on a slightly different potential function, namely,

$$\mathcal{U}_{k+1} := \|z^{k+1} - z^\star\|_{\Gamma^{-1}}^2 + A_{k+1}\|s^k - S_{z^k}(z^k;\bar{z}^k)\|_\Gamma^2 + B_{k+1}\|z^{k+1} - z^k\|_\Gamma^2$$

where $(A_k)_{k\in\mathbb{N}}$ and $(B_k)_{k\in\mathbb{N}}$ are positive scalar parameters to be identified.

We start by writing out one step of the update

$$\|z^{k+1} - z^\star\|_{\Gamma^{-1}}^2 = \|z^k - z^\star\|_{\Gamma^{-1}}^2 - 2\alpha_k\langle h^k - \hat{\hat{H}}_k, z^k - z^\star\rangle_\Gamma + \alpha_k^2\|h^k - \hat{\hat{H}}_k\|_\Gamma^2 \tag{E.9}$$

$$= \|z^k - z^\star\|_{\Gamma^{-1}}^2 - 2\alpha_k\langle s^k - \hat{\hat{S}}_k, z^k - z^\star\rangle_\Gamma + \alpha_k^2\|s^k - \hat{\hat{S}}_k\|_\Gamma^2 \tag{E.10}$$

In the algorithm, s^k estimates $S_{z^k}(z^k;\bar{z}^k)$. Recall $M_{z^k}(z) := F(z) - Q_{z^k}(z)$ and $\hat{M}_{z^k}(z,\xi) := \hat{F}(z,\xi) - \hat{Q}_{z^k}(z,\xi)$. Let us quantify how good this estimation is. We will make use of the careful choice of the bias-correction term to shift the noise index by 1 in the second inequality.

$$\begin{aligned}
s^k - S_{z^k}(z^k;\bar{z}^k) &= M_{z^k}(z^k) + Q_{z^k}(\bar{z}^k) - \hat{M}_{z^k}(z^k,\xi_k) - \hat{Q}_{z^k}(\bar{z}^k,\xi_k') \\
&\quad + (1-\alpha_{k-1})(h^{k-1} - \Gamma^{-1}z^{k-1} + \hat{M}_{z^{k-1}}(z^{k-1},\xi_k) - \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_{k-1}') + \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')) \\
&= M_{z^k}(z^k) + Q_{z^k}(\bar{z}^k) - \hat{M}_{z^k}(z^k,\xi_k) - \hat{Q}_{z^k}(\bar{z}^k,\xi_k') \\
&\quad + (1-\alpha_{k-1})(s^{k-1} + \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\bar{\xi}_k') - \Gamma^{-1}z^{k-1} + \hat{M}_{z^{k-1}}(z^{k-1},\xi_k)) \\
&= M_{z^k}(z^k) + Q_{z^k}(\bar{z}^k) - \hat{M}_{z^k}(z^k,\xi_k) - \hat{Q}_{z^k}(\bar{z}^k,\xi_k') + (1-\alpha_{k-1})(s^{k-1} - S_{z^{k-1}}(z^{k-1};\bar{z}^{k-1})) \\
&\quad + (1-\alpha_{k-1})(\hat{M}_{z^{k-1}}(z^{k-1},\xi_k) - M_{z^{k-1}}(z^{k-1}) + \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\bar{\xi}_k') - Q_{z^{k-1}}(\bar{z}^{k-1}))
\end{aligned}$$

Using the shorthand notation

$$\begin{aligned}
\tilde{s}^k &:= s^k - S_{z^k}(z^k;\bar{z}^k), \\
\tilde{Q}_{z^k}(\bar{z}^k,\xi_k') &:= Q_{z^k}(\bar{z}^k) - \hat{Q}_{z^k}(\bar{z}^k,\xi_k'), \\
\tilde{M}(z^k,\xi_k) &:= M_{z^k}(z^k) - \hat{M}_{z^k}(z^k,\xi_k),
\end{aligned}$$

it follows that

$$\begin{aligned}
\|\tilde{s}^k\|_\Gamma^2 &= (1-\alpha_{k-1})^2\|\tilde{s}^{k-1}\|_\Gamma^2 + \|\tilde{M}(z^k,\xi_k) + \tilde{Q}_{z^k}(\bar{z}^k,\xi_k') - (1-\alpha_{k-1})(\tilde{M}(z^{k-1},\xi_k) + \tilde{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')\|_\Gamma^2 \\
&\quad + 2(1-\alpha_{k-1})\langle \tilde{s}^{k-1}, \tilde{M}(z^k,\xi_k) + \tilde{Q}_{z^k}(\bar{z}^k,\xi_k') - (1-\alpha_{k-1})(\tilde{M}(z^{k-1},\xi_k) + \tilde{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k'))\rangle_\Gamma
\end{aligned} \tag{E.11}$$

In the scalar product, the left term is known when z^k is known. Moreover, since $\mathbb{E}[\cdot \mid \mathcal{F}_k] = \mathbb{E}[\mathbb{E}[\cdot \mid \mathcal{F}_k'] \mid \mathcal{F}_k]$, owing to $\mathcal{F}_k \subset \mathcal{F}_k'$, we have

$$\mathbb{E}\Big[\tilde{M}(z^k,\xi_k) + \tilde{Q}_{z^k}(\bar{z}^k,\xi_k') - (1-\alpha_{k-1})(\tilde{M}(z^{k-1},\xi_k) + \tilde{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')) \mid \mathcal{F}_k\Big] = \mathbb{E}\Big[\tilde{Q}_{z^k}(\bar{z}^k,\xi_k') \mid \mathcal{F}_k\Big] = 0,$$

where we use Assumption II*(i)* through Lemma E.2.

Since the second moment is larger than the variance we have

$$\begin{aligned}
&\mathbb{E}\Big[\|\tilde{M}(z^k,\xi_k) + \tilde{Q}_{z^k}(\bar{z}^k,\xi_k') - \tilde{M}(z^{k-1},\xi_k) - \tilde{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k\Big] \le \\
&\qquad \mathbb{E}\Big[\|\hat{M}_{z^k}(z^k,\xi_k) - \hat{M}_{z^{k-1}}(z^{k-1},\xi_k) + \hat{Q}_{z^k}(\bar{z}^k,\xi_k') - \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k\Big]
\end{aligned} \tag{E.12}$$

Using the Young inequality it follows from (E.11), (E.12) that

$$\begin{aligned}
\mathbb{E}[\|\tilde{s}^k\|_\Gamma^2 \mid \mathcal{F}_k] &\le (1-\alpha_{k-1})^2\|\tilde{s}^{k-1}\|_\Gamma^2 + 2\alpha_{k-1}^2\mathbb{E}[\|\tilde{M}(z^k,\xi_k) + \tilde{Q}_{z^k}(\bar{z}^k,\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k] \\
&\quad + 2(1-\alpha_{k-1})^2\mathbb{E}[\|\tilde{M}(z^k,\xi_k) + \tilde{Q}_{z^k}(\bar{z}^k,\xi_k') - \tilde{M}(z^{k-1},\xi_k) - \tilde{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k] \\
&\le (1-\alpha_{k-1})^2\|\tilde{s}^{k-1}\|_\Gamma^2 + 2\alpha_{k-1}^2\mathbb{E}[\|M_{z^k}(z^k) - \hat{M}_{z^k}(z^k,\xi_k) + Q_{z^k}(\bar{z}^k) - \hat{Q}_{z^k}(\bar{z}^k,\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k] \\
&\quad + \mathbb{E}[2(1-\alpha_{k-1})^2\|\hat{M}_{z^k}(z^k,\xi_k) - \hat{M}_{z^{k-1}}(z^{k-1},\xi_k) + \hat{Q}_{z^k}(\bar{z}^k,\xi_k') - \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k] \\
&\le (1-\alpha_{k-1})^2\|\tilde{s}^{k-1}\|_\Gamma^2 + 2\alpha_{k-1}^2\sigma_F^2 \\
&\quad + \mathbb{E}[2(1-\alpha_{k-1})^2\|\hat{M}_{z^k}(z^k,\xi_k) - \hat{M}_{z^{k-1}}(z^{k-1},\xi_k) + \hat{Q}_{z^k}(\bar{z}^k,\xi_k') - \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k]
\end{aligned} \tag{E.13}$$

where the last inequality follows from Assumptions II*(i)* and II*(ii)* through Lemma E.2.

To bound the last term of (E.13) we use the particular choice of Q_u,

$$\begin{aligned}
&\hat{M}_{z^k}(z^k,\xi_k) - \hat{M}_{z^{k-1}}(z^{k-1},\xi_k) + \hat{Q}_{z^k}(\bar{z}^k,\xi_k') - \hat{Q}_{z^{k-1}}(\bar{z}^{k-1},\xi_k') \\
&\quad = \begin{pmatrix} \nabla_x\hat{\varphi}(z^k,\xi) - \nabla_x\hat{\varphi}(z^{k-1},\xi) \\ (1-\theta)(\nabla_y\hat{\varphi}(z^{k-1}\xi_k) - \nabla_y\hat{\varphi}(z^k,\xi_k)) - \theta(\nabla_y\hat{\varphi}(\bar{x}^k,y^k,\xi_k') - \nabla_y\hat{\varphi}(\bar{x}^k,y^k,\xi_k') \end{pmatrix}.
\end{aligned} \tag{E.14}$$

So Requirement IV.A4 applies, leading to the following bound

$$\begin{aligned}
&\mathbb{E}[\|M_{z^k}(z^k) - \hat{M}_{z^k}(z^k,\xi_k) + Q_{z^k}(\bar{z}^k) - \hat{Q}_{z^k}(\bar{z}^k,\xi_k')\|_\Gamma^2 \mid \mathcal{F}_k] \\
&\qquad \le \|z^k - z^{k-1}\|_{D_{\widehat{xz}}+D_{\widehat{yz}}}^2 + \|y^k - y^{k-1}\|_{D_{\widehat{yy}}}^2 + \|\bar{x}^k - \bar{x}^{k-1}\|_{D_{\widehat{yx}}}^2 \\
&\qquad \le \hat{c}_1\|z^k - z^{k-1}\|_{\Gamma^{-1}}^2 + \hat{c}_2\|\bar{x}^k - \bar{x}^{k-1}\|_{\Gamma^{-1}}^2
\end{aligned} \tag{E.15}$$

where $\hat{c}_1 := \|\Gamma(D_{\widehat{xz}} + D_{\widehat{yz}}) + \Gamma_2 D_{\widehat{yy}}\|$ and $\hat{c}_2 := \|\Gamma_1 D_{\widehat{yx}}\|$. Using (E.15) in (E.13) yields,

$$\mathbb{E}[\|\tilde{s}^k\|_\Gamma^2 \mid \mathcal{F}_k] \le (1-\alpha_{k-1})^2\|\tilde{s}^{k-1}\|_\Gamma^2 + 2\alpha_{k-1}^2\sigma_F^2 + 2(1-\alpha_{k-1})^2\Big(\hat{c}_1\|z^k - z^{k-1}\|_{\Gamma^{-1}}^2 + \hat{c}_2\|\bar{x}^k - \bar{x}^{k-1}\|_{\Gamma_1^{-1}}^2\Big). \tag{E.16}$$

By monotonicity of A and since $s^k \in \Gamma^{-1}\bar{z}^k + A\bar{z}^k$ we have

$$0 \le \langle s^k - \Gamma^{-1}\bar{z}^k - s^{k-1} + \Gamma^{-1}\bar{z}^{k-1}, \bar{z}^k - \bar{z}^{k-1}\rangle = -\|\bar{z}^k - \bar{z}^{k-1}\|^2_{\Gamma^{-1}} + \langle s^k - s^{k-1}, \bar{z}^k - \bar{z}^{k-1}\rangle.$$

Therefore, using the Cauchy–Schwarz inequality

$$\|\bar{z}^k - \bar{z}^{k-1}\|^2_{\Gamma^{-1}} \le \langle s^k - s^{k-1}, \bar{z}^k - \bar{z}^{k-1}\rangle \le \|\bar{z}^k - \bar{z}^{k-1}\|_{\Gamma^{-1}}\|s^k - s^{k-1}\|_{\Gamma}. \tag{E.17}$$

To majorize $\|\bar{x}^k - \bar{x}^{k-1}\|_{\Gamma_1^{-1}}$ in (E.16) let s_x^k be the primal components of s^k in what follows. Due to nonexpansiveness derived in (E.17),

$$\begin{aligned}
\|\bar{x}^k - \bar{x}^{k-1}\|_{\Gamma_1^{-1}} &\le \|s_x^k - s_x^{k-1}\|_{\Gamma_1} \\
&= \|\Gamma_1^{-1}x^k - \nabla_x\hat{\phi}(z^k,\xi_k) + (1-\alpha_{k-1})(\Gamma_1^{-1}(x^{k-1} - x^{k-1}) + \hat{\nabla}_x\hat{\phi}(z^{k-1},\xi_k)) - s_x^{k-1}\|_{\Gamma_1} \\
&\le (1-\alpha_{k-1})\|x^k - x^{k-1}\|_{\Gamma_1^{-1}} + (1-\alpha_{k-1})\|\nabla_x\hat{\phi}(z^k,\xi_k) - \nabla_x\hat{\phi}(z^{k-1},\xi_k)\|_{\Gamma_1} \\
&\qquad + \alpha_{k-1}\|\Gamma_1^{-1}x^k - \nabla_x\hat{\phi}(z^k,\xi_k) - s_x^{k-1}\|_{\Gamma_1} \\
\text{(Requirement IV.A4)} &\le (1-\alpha_{k-1})\|x^k - x^{k-1}\|_{\Gamma_1^{-1}} + (1-\alpha_{k-1})\|z^k - z^{k-1}\|_{D_{\widehat{xz}}} \\
&\qquad + \alpha_{k-1}\|s_x^k - s_x^{k-1}\|_{\Gamma_1^{-1}} + \alpha_{k-1}(1-\alpha_{k-1})\|\Gamma_1^{-1}x^{k-1} - \nabla_x\hat{\phi}(z^k,\xi_k) - s_x^{k-1}\|_{\Gamma_1}
\end{aligned}$$

Hence, rearranging the $\alpha_{k-1}\|s_x^k - s_x^{k-1}\|$ term, we get

$$\begin{aligned}
\mathbb{E}[\|\bar{x}^k - \bar{x}^{k-1}\|^2 \mid \mathcal{F}_k] &\le \mathbb{E}[\|s_x^k - s_x^{k-1}\|^2 \mid \mathcal{F}_k] \\
&\le 2(1+\hat{c}_3)\|z^k - z^{k-1}\|^2_{\Gamma_1^{-1}} + 2\alpha_{k-1}^2\mathbb{E}[\|\Gamma^{-1}x^{k-1} - \nabla_x\hat{\phi}(z^{k-1},\xi_k) - s_x^{k-1}\|^2_{\Gamma_1} \mid \mathcal{F}_k] \\
\text{Assumptions II(i) and II(ii)} &\le 2(1+\hat{c}_3)\|z^k - z^{k-1}\|^2_{\Gamma_1^{-1}} + 2\alpha_{k-1}^2\mathbb{E}[\|\Gamma^{-1}x^{k-1} - \nabla_x\phi(z^{k-1}) - s_x^{k-1}\|^2_{\Gamma_1} \mid \mathcal{F}_k] + 2\alpha_{k-1}^2\sigma_F^2 \\
&\le 2(1+\hat{c}_3)\|z^k - z^{k-1}\|^2_{\Gamma_1^{-1}} + 2\alpha_{k-1}^2\mathbb{E}[\|s^{k-1} - S_{z^{k-1}}(z^{k-1};\bar{z}^{k-1})\|^2_{\Gamma} \mid \mathcal{F}_k] + 2\alpha_{k-1}^2\sigma_F^2
\end{aligned}$$

where $\hat{c}_3 := \|D_{\widehat{xz}}\Gamma\|$.

We finally obtain

$$\mathbb{E}[\|\bar{z}^k - \bar{z}^{k-1}\|^2_{\Gamma^{-1}} \mid \mathcal{F}_k] \le 2(1+\hat{c}_3)\|z^k - z^{k-1}\|^2_{\Gamma^{-1}} + 2\alpha_{k-1}^2\mathbb{E}[\|s^{k-1} - S_{z^{k-1}}(z^{k-1};\bar{z}^{k-1})\|^2_{\Gamma} \mid \mathcal{F}_k] + 2\alpha_{k-1}^2\sigma_F. \tag{E.18}$$

Introducing (E.18) into (E.16) yields

$$\begin{aligned}
\mathbb{E}[\|s^k - S_{z^k}(z^k;\bar{z}^k)\|^2_{\Gamma} \mid \mathcal{F}_k] &\le (1-\alpha_{k-1})^2(1+4\hat{c}_2\alpha_{k-1}^2)\|s^{k-1} - S_{z^{k-1}}(z^{k-1};\bar{z}^{k-1})\|^2_{\Gamma} \\
&\quad + 2(1-\alpha_{k-1})^2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3))\|z^k - z^{k-1}\|^2_{\Gamma^{-1}} \\
&\quad + 2\alpha_{k-1}^2(1+(1-\alpha_{k-1})^2 2\hat{c}_2)\sigma_F^2.
\end{aligned} \tag{E.19}$$

We continue with the inner term in (E.10).

$$\begin{aligned}
&-\mathbb{E}[\langle s^k - \hat{\bar{S}}_k, z^k - z^\star\rangle_\Gamma \mid \mathcal{F}_k] \\
&\quad = \langle -s^k + S_{z^k}(\bar{z}^k), z^k - z^\star\rangle_\Gamma \\
&\quad = \langle -s^k + S_{z^k}(\bar{z}^k), z^k - \bar{z}^k\rangle_\Gamma + \langle -s^k + S_{z^k}(\bar{z}^k), \bar{z}^k - z^\star\rangle_\Gamma \\
&\quad = \langle -s^k + S_{z^k}(z^k;\bar{z}^k), z^k - \bar{z}^k\rangle_\Gamma - \langle S_{z^k}(z^k;\bar{z}^k) - S_{z^k}(\bar{z}^k), z^k - \bar{z}^k\rangle_\Gamma + \langle -s^k + S_{z^k}(\bar{z}^k), \bar{z}^k - z^\star\rangle_\Gamma \\
&\quad = \langle -s^k + S_{z^k}(z^k;\bar{z}^k), z^k - \bar{z}^k\rangle_\Gamma - \langle H_{z^k}(z^k) - H_{z^k}(\bar{z}^k), z^k - \bar{z}^k\rangle_\Gamma + \langle -s^k + S_{z^k}(\bar{z}^k), \bar{z}^k - z^\star\rangle_\Gamma
\end{aligned}$$

By definition of $\bar{z}^k$ in (8.4b), we have $s^k = h^k - \hat{Q}_{z^k}(\bar{z}^k,\xi_k') \in \Gamma^{-1}\bar{z}^k + A(\bar{z}^k)$, so that $s^k - S_{z^k}(\bar{z}^k) \in F(\bar{z}^k) + A(\bar{z}^k)$. Hence, using the weak Minty assumption,

$$\langle s^k - S_{z^k}(\bar{z}^k), \bar{z}^k - z^\star\rangle_\Gamma \ge \rho\|s^k - S_{z^k}(\bar{z}^k)\|^2 . \tag{E.20}$$

Using also cocoercivity derived in (E.7), this leads to the following inequality, true for any $\varepsilon_k > 0$:

$$\begin{aligned}
-\mathbb{E}[\langle s^k - \hat{\bar{S}}_k, z^k - z^\star\rangle \mid \mathcal{F}_k] &\le \tfrac{\varepsilon_k}{2}\|s^k - S_{z^k}(z^k;\bar{z}^k)\|^2_{\Gamma} + \tfrac{1}{2\varepsilon_k}\|\bar{z}^k - z^k\|^2_{\Gamma^{-1}} \\
&\quad - \tfrac{1}{2}\|H_{z^k}(z^k) - H_{z^k}(\bar{z}^k)\|^2_{\Gamma} - \rho\|s^k - S_{z^k}(\bar{z}^k)\|^2 .
\end{aligned}$$

To majorize the term $\|\bar{z}^k - z^k\|^2_{\Gamma^{-1}}$, we may use the fact that

$$\begin{aligned}
\|H_{z^k}(z) - H_{z^k}(z')\|^2_{\Gamma} &= \|\Gamma^{-1}(z - z') - M_{z^k}(z) + M_{z^k}(z')\|^2_{\Gamma} \\
&\ge \|z - z'\|^2_{\Gamma^{-1}} - \|M_{z^k}(z) - M_{z^k}(z')\|^2_{\Gamma} \\
\text{Requirement IV.A2} &\ge (1 - L_M^2)\|z - z'\|^2_{\Gamma^{-1}} .
\end{aligned}$$

where $L_M^2 \coloneqq \mathbf{max}\{\|D_{xx}\Gamma_1\| + \|D_{xy}\Gamma_2\|, \|D_{yx}\Gamma_1\| + \|D_{yy}\Gamma_2\|\}$. Hence,

$$\begin{aligned}-\mathbb{E}[\langle s^k - \hat{\bar{S}}_k, z^k - z^\star\rangle_\Gamma \mid \mathcal{F}_k] \leq \tfrac{\varepsilon_k}{2}\|s^k - S_{z^k}(z^k;\bar{z}^k)\|_\Gamma^2 \\ + \Big(\tfrac{1}{2\varepsilon_k(1-L_M^2)} - \tfrac{1}{2}\Big)\|H_{z^k}(z^k) - H_{z^k}(\bar{z}^k)\|_\Gamma^2 - \rho\|s^k - S_{z^k}(\bar{z}^k)\|^2 .\end{aligned} \tag{E.21}$$

The third term in (E.10) is bounded by

$$\begin{aligned}\alpha_k^2\mathbb{E}[\|s^k - \hat{\bar{S}}_k\|_\Gamma^2 \mid \mathcal{F}_k] &= \alpha_k^2\|s^k - S_{z^k}(\bar{z}^k)\|_\Gamma^2 + \alpha_k^2\mathbb{E}[\|F(\bar{z}^k) - \hat{F}(\bar{z}^k,\bar{\xi}_k)\|_\Gamma^2 \mid \mathcal{F}_k] \\ &\leq \alpha_k^2\|s^k - S_{z^k}(\bar{z}^k)\|_\Gamma^2 + \alpha_k^2\sigma_F^2\end{aligned} \tag{E.22}$$

where we have used Assumption II*(ii)*.

Combined with the update rule, (E.22) can also be used to bound the difference of iterates

$$\mathbb{E}[\|z^{k+1} - z^k\|_{\Gamma^{-1}}^2 \mid \mathcal{F}_k] = \mathbb{E}[\alpha_k^2\|s^k - \hat{\bar{S}}_k\|_\Gamma^2 \mid \mathcal{F}_k] \leq \alpha_k^2\|s^k - S_{z^k}(\bar{z}^k)\|_\Gamma^2 + \alpha_k^2\sigma_F^2 \tag{E.23}$$

Using (E.10), (E.21), (E.22), (E.23) and that $-\rho\|s^k - S_{z^k}(\bar{z}^k)\|^2 \leq -\frac{\rho}{\bar{\gamma}}\|s^k - S_{z^k}(\bar{z}^k)\|_\Gamma^2$ with $\bar{\gamma}$ denoting the smallest eigenvalue of Γ we have,

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] &\leq \|z^k - z^\star\|_{\Gamma^{-1}}^2 + (A_{k+1} + \alpha_k\varepsilon_k)\|s^k - S_{z^k}(z^k;\bar{z}^k)\|_\Gamma^2 + \alpha_k\Big(\frac{1}{\varepsilon_k(1-L_M^2)} - 1\Big)\|H_{z^k}(z^k) - H_{z^k}(\bar{z}^k)\|_\Gamma^2 \\ &\quad + \alpha_k(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k B_{k+1})\|s^k - S_{z^k}(\bar{z}^k)\|_\Gamma^2 + \alpha_k^2(1+B_{k+1})\sigma_F^2 \\ &\leq \|z^k - z^\star\|_{\Gamma^{-1}}^2 + (A_{k+1} + \alpha_k(\varepsilon_k + (1+1/b)(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k B_{k+1})))\|s^k - S_{z^k}(z^k;\bar{z}^k)\|_\Gamma^2 \\ &\quad + \alpha_k\Big(\frac{1}{\varepsilon_k(1-L_M^2)} - 1 + (1+b)(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k B_{k+1})\Big)\|H_{z^k}(z^k) - H_{z^k}(\bar{z}^k)\|_\Gamma^2 \\ &\quad + \alpha_k^2(1+B_{k+1})\sigma_F^2,\end{aligned} \tag{E.24}$$

where the last inequality follows from Young's inequality with positive b.

By defining

$$\begin{aligned}X_k^1 &\coloneqq A_{k+1} + \alpha_k(\varepsilon_k + (1+1/b)(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k B_{k+1})) \\ X_k^2 &\coloneqq \alpha_k\Big(\frac{1}{\varepsilon_k(1-L_M^2)} - 1 + (1+b)(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k B_{k+1})\Big)\end{aligned} \tag{E.25}$$

and applying (E.19), we finally obtain

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k &\leq X_k^2\|H_{z^k}(z^k) - H_{z^k}(\bar{z}^k)\|_\Gamma^2 \\ &\quad + (X_k^1(1-\alpha_{k-1})^2(1 + 4\hat{c}_2\alpha_{k-1}^2) - A_k)\|s^{k-1} - S_{z^{k-1}}(z^{k-1};\bar{z}^{k-1})\|_\Gamma^2 \\ &\quad + (2X_k^1(1-\alpha_{k-1})^2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)) - B_k)\|z^k - z^{k-1}\|_{\Gamma^{-1}}^2 \\ &\quad + 2X_k^1\alpha_{k-1}^2(1 + (1-\alpha_{k-1})^2 2\hat{c}_2)\sigma_F^2 + \alpha_k^2(1+B_{k+1})\sigma_F^2,\end{aligned} \tag{E.26}$$

If $A_k \geq X_k^1(1-\alpha_{k-1})^2(1+4\hat{c}_2\alpha_{k-1}^2)$, then it suffice to pick B_k as

$$2X_k^1(1-\alpha_{k-1})^2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)) = \tfrac{2(\hat{c}_1+2\hat{c}_2(1+\hat{c}_3))A_k}{1+4\hat{c}_2\alpha_{k-1}^2} \leq 2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3))A_k =: B_k. \tag{E.27}$$

To get a recursion, we then only require the following conditions

$$X_k^1(1-\alpha_{k-1})^2(1+4\hat{c}_2\alpha_{k-1}^2) \leq A_k \quad \text{and} \quad X_k^2 < 0. \tag{E.28}$$

Set $A_k = A$, $\varepsilon_k = \varepsilon$. For the first inequality of (E.28), since $(1-\alpha_k)^2 \leq (1-\alpha_k)$, the terms involving A are bounded as

$$\begin{aligned}&(1-\alpha_k)^2(1+4\hat{c}_2\alpha_k^2)(1+2\alpha_k^2(1+1/b)(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)))A - A \\ &\quad \leq (1-\alpha_k)(1+4\hat{c}_2\alpha_k^2)(1+2\alpha_k^2(1+1/b)(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)))A - A \\ &\quad = -\alpha_k A + (1-\alpha_k)(4\hat{c}_2\alpha_k^2)A + (1-\alpha_k)(1+4\hat{c}_2\alpha_k^2)(2\alpha_k^2(1+1/b)(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)))A \\ &\quad \leq -\alpha_k\Big(1 - 4\hat{c}_2\alpha_0 - (1+4\hat{c}_2\alpha_0^2)(2\alpha_0(1+1/b)(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)))\Big)A\end{aligned} \tag{E.29}$$

where the last inequality follows from $(1-\alpha_k) \leq 1$ and $\alpha_k \leq \alpha_0$. Thus to satisfy the first inequality of (E.28) it suffice to pick

$$A \geq \frac{\varepsilon + (1+1/b)(\alpha_0 - \frac{2\rho}{\bar{\gamma}})}{\nu} \tag{E.30}$$

where $\nu > 0$ with ν as defined in (E.1).

The second equality of (E.28) is satisfied owing to (E.1).

The noise term in (E.26) can be made independent of k by using $\alpha_k \leq \alpha_0$,

$$\begin{aligned}
&2X_k^1(1+(1-\alpha_k)^2 2\hat{c}_2) + 1 + B_{k+1} \\
&\quad = 2(A + \alpha_k(\varepsilon + (1+1/b)(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k 2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)A)))(1+(1-\alpha_k)^2 2\hat{c}_2) \\
&\qquad + 1 + 2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)A \\
&\quad \leq 2(A + \alpha_0(\varepsilon + (1+1/b)(\alpha_0 - 2\rho/\bar{\gamma} + \alpha_0 2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)A)))(1+2\hat{c}_2) \\
&\qquad + 1 + 2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)A =: C.
\end{aligned} \tag{E.31}$$

Thus, it follows from (E.26) that

$$\begin{aligned}
&\mathbb{E}[\mathcal{U}_{k+1} \mid \mathcal{F}_k] - \mathcal{U}_k \\
&\quad \leq \alpha_k\Big(\frac{1}{\varepsilon(1-L_M^2)} - 1 + (1+b)(\alpha_k - 2\rho/\bar{\gamma} + \alpha_k 2(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3)A)\Big)\|H_{z^k}(z^k) - H_{z^k}(\bar{z}^k)\|_\Gamma^2 \\
&\qquad + \alpha_k^2 C\sigma_F^2.
\end{aligned} \tag{E.32}$$

The result is obtained by total expectation and summing the above inequality while noting that the initial iterate were set as $z^{-1} = z^0$. □

Proof of Theorem 8.2. The theorem is a specialization of Theorem E.1 for a particular a choice of b and ε. The second requirement of (E.1) can be rewritten as,

$$b > \frac{\alpha_0 d}{1 - 4\alpha_0\hat{c}_2 - \alpha_0 d}, \tag{E.33}$$

where $d = 2(1+4\hat{c}_2\alpha_0^2)(\hat{c}_1 + 2\hat{c}_2(1+\hat{c}_3))$. This condition is satisfied by $b = \frac{2\alpha_0 d}{1-\alpha_0(4\hat{c}_2+d)}$. We substitute in the choice of b, $\varepsilon = \frac{1}{(1-L_M^2)\sqrt{\alpha_0}}$ and denotes $\eta := A$.

The proof is complete by converting the weighted sum into an expectation over a sampled iterate in the style of Ghadimi & Lan (2013). □

E.1 Explanation of bias-correction term

Consider the naive analysis which would track h^k. By the definition of $\bar{z}^k$ in (8.4b) and $H_k(\bar{z}^k)$ we would have $h^k - H_k(\bar{z}^k) + P_k(\bar{z}^k) - \hat{P}_k(\bar{z}^k, \bar{\xi}_k) \in F(\bar{z}^k) + A(\bar{z}^k)$. Hence, assuming zero mean and using the weak MVI from Assumption I*(iii)*,

$$\begin{aligned}
\mathbb{E}[\langle h^k - H_k(\bar{z}^k), \bar{z}^k - z^\star\rangle \mid \mathcal{F}_k'] &= \mathbb{E}[\langle h^k - H_k(\bar{z}^k) + P_k(\bar{z}^k) - \hat{P}_k(\bar{z}^k, \xi_k'), \bar{z}^k - z^\star\rangle \mid \mathcal{F}_k'] \\
&\geq \mathbb{E}[\rho\|h^k - H_k(\bar{z}^k) + P_k(\bar{z}^k) - \hat{P}_k(\bar{z}^k, \xi_k')\|^2 \mid \mathcal{F}_k'] \,.
\end{aligned} \tag{E.34}$$

To proceed we could apply Young's inequality, but this would produce a noise term, which would propagate to the descent inequality in (E.26) with a α_k factor in front. To show convergence we would instead need a smaller factor of α_k^2.

To avoid this error term entirely we instead do a change of variables with $s^k := h^k - \hat{P}_{z^k}(\bar{z}^k, \xi_k')$ such that,

$$\begin{aligned}
h^k \in \hat{P}_{z^k}(\bar{z}^k, \xi_k') + A\bar{z}^k &\Leftrightarrow h^k - \hat{P}_{z^k}(\bar{z}^k, \xi_k') \in A\bar{z}^k \\
&\Leftrightarrow s^k \in A\bar{z}^k.
\end{aligned} \tag{E.35}$$

This make application of Assumption I*(iii)* unproblematic, but affects the choice of the bias-correction term, since the analysis will now apply to s^k. If we instead of the careful choice of h^k

in (8.4a) had made the choice

$$h^k = \hat{P}_{z^k}(z^k, \xi_k) - \hat{F}(z^k, \xi_k) + (1-\alpha_k)(h^{k-1} - \hat{P}_{z^{k-1}}(z^{k-1}, \xi_k) + \hat{F}(z^{k-1}, \xi_k)) \tag{E.36}$$

then

$$s^k = \hat{P}_{z^k}(z^k, \xi_k) - \hat{F}(z^k, \xi_k) - \hat{P}_{z^k}(\bar{z}^k, \xi'_k) + (1-\alpha_k)(s^{k-1} - \hat{P}_{z^{k-1}}(z^{k-1}, \xi_k) + \hat{F}(z^{k-1}, \xi_k) - \hat{P}_{z^{k-1}}(\bar{z}^{k-1}, \xi'_{k-1})). \tag{E.37}$$

Notice how the latter term is evaluated under ξ'_{k-1} instead of ξ'_k. The choice in (8.4a) resolves this issue.

F Negative weak Minty variational inequality

In this section we consider the problem of finding a zero of the single-valued operator F (with the set-valued operator $A \equiv 0$). Observe that the WMVI assumption Assumption I*(iii)* $\langle Fz, z - z^\star \rangle \geq \rho \|Fz\|^2$, for all $z \in \mathbb{R}^n$, is not symmetric and one may instead consider that the assumption holds for $-F$. As we will see below this simple observation leads to nontrivial problem classes extending the reach of extragradient-type methods both in the deterministic and stochastic settings.

Assumption VII (negative weak MVI)**.** *There exists a nonempty set $\mathcal{S}^\star \subseteq \mathbf{zer}\, T$ such that for all $z^\star \in \mathcal{S}^\star$ and some $\bar{\rho} \in (-\infty, 1/2L)$*

$$\langle Fz, z - z^\star \rangle \leq \bar{\rho} \|Fz\|^2, \quad \textit{for all } z \in \mathbb{R}^n. \tag{F.1}$$

Under this assumption the algorithm of Pethick et al. (2022) leads to the following modified iterates:

$$\bar{z}^k = z^k + \gamma_k F z^k, \tag{F.2}$$

$$z^{k+1} = z^k + \lambda_k \alpha_k (H_k \bar{z}^k - H_k z^k) = z^k + \lambda_k \alpha_k \gamma_k F \bar{z}^k, \quad \text{where} \quad H_k := \mathrm{id} + \gamma_k F \tag{F.3}$$

We next consider the lower bound example of (Pethick et al., 2022, Ex. 5) to show that despite the condition for weak MVI being violated for b smaller than a certain threshold, the negative weak MVI in Assumption VII holds for any negative b and thus the extragradient method applied to $-F$ is guaranteed to converge.

Example 1. Consider (Pethick et al., 2022, Ex. 5)

$$\underset{x \in \mathbb{R}}{\textbf{minimize}}\ \underset{y \in \mathbb{R}}{\textbf{maximize}}\ f(x, y) := axy + \frac{b}{2}(x^2 - y^2), \tag{F.4}$$

where $b < 0$ and $a > 0$. The associated F is a linear mapping. For a linear mapping M, Assumption VII holds if

$$\tfrac{1}{2}(M + M^\top) - \bar{\rho} M^\top M \preceq 0, \qquad \bar{\rho} \in (-\infty, 1/2L)$$

While Assumption I*(iii)* holds if

$$\tfrac{1}{2}(M + M^\top) - \rho M^\top M \succeq 0, \qquad \rho \in (-1/2L, \infty).$$

For this example $L = \sqrt{a^2 + b^2}$ and

$$F(z) = \overbrace{(bx + ay, -ax + by)}^{=:Mz}.$$

Since M is a bisymmetric linear mapping, $M^\top M = (a^2 + b^2)\mathrm{I}$ which according to the above characterizations implies

$$\rho \in (-\tfrac{1}{2L}, \tfrac{b}{a^2+b^2}], \qquad \bar{\rho} \in [\tfrac{b}{a^2+b^2}, \tfrac{1}{2L}).$$

The range for ρ is nonempty if $b > -\frac{a}{\sqrt{3}}$ while this is not an issue for $\bar{\rho}$ which allows any negative b.

We complete this section with a corollary to Theorem 6.3 when replacing weak MVI assumption with Assumption VII.

Corollary F.1. *Suppose that Assumptions I(i) to III and VII hold. Let $(z^k)_{k \in \mathbb{N}}$ denote the sequence generated by algorithm 1 applied to $-F$. Then, the claims of Theorem 6.3 hold true.*

G Experiments

G.1 Synthetic example

Example 2 (Unconstrained quadratic game (Pethick et al., 2022, Ex. 5))**.** Consider,

$$\underset{x \in \mathbb{R}}{\textbf{minimize}}\ \underset{y \in \mathbb{R}}{\textbf{maximize}}\ \phi(x, y) := axy + \frac{b}{2}x^2 - \frac{b}{2}y^2, \tag{G.1}$$

where $a \in \mathbb{R}_+$ and $b \in \mathbb{R}$.

The problem constants in Example 2 can easily be computed as $\rho = \frac{b}{a^2+b^2}$ and $L = \sqrt{a^2+b^2}$. We can rewrite Example 2 in terms of L and ρ by choosing $a = \sqrt{L^2 - L^4\rho^2}$ and $b = L^2\rho$.

Example 3 (Constrained minimax (Pethick et al., 2022, Ex. 4))**.** Consider

$$\underset{|x|\leq 4/3}{\textbf{minimize}}\ \underset{|y|\leq 4/3}{\textbf{maximize}}\ \phi(x,y) := xy + \psi(x) - \psi(y), \tag{GlobalForsaken}$$

where $\psi(z) = \frac{2z^6}{21} - \frac{z^4}{3} + \frac{z^2}{3}$.

In both Example 2 and Example 3 the operator F is defined as $Fz = (\nabla_x\phi(x,y), -\nabla_y\phi(x,y))$.

We compare BCSEG+ against (SEG) (with $\alpha_0 = \beta_0$) and (EG+) in both an unconstrained setting and a constrained setting introduced in Pethick et al. (2022). In the unconstrained example we control all problem constant and set $\rho = -1/10L_F$ if not otherwise noted, while the constrained example is a specific minimax problem where $\rho > -1/2L_F$ holds within the constraint set for a Lipschitz constant L_F restricted to the same constraint set. To simulate a stochastic setting in both example, we consider additive Gaussian noise, i.e. $\hat{F}(z,\xi) = Fz + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2 I)$. When not stated otherwise, we choose $\sigma = 0.1$ and initialize with $z^0 = \mathbf{1}$.

The default configuration is $\gamma = 1/2L_F$, $\alpha_k = 1/18 \cdot \sqrt{k/100+1}$ and $\beta_k = \beta_0\alpha_k/\alpha_0$.

G.2 Additional algorithmic details

For the constrained setting in Figure 1, we consider two extensions of (SEG+). One variant uses a single application of the resolvent as suggested by Pethick et al. (2022),

$$\begin{aligned} \bar{z}^k &= (\mathrm{id} + \gamma A)^{-1}(z^k - \gamma\hat{F}(z^k, \xi_k)) \quad \text{with} \quad \xi_k \sim \mathcal{P} \\ z^{k+1} &= z^k + \alpha_k\Big((\bar{z}^k - z^k) - \gamma(\hat{F}(\bar{z}^k, \bar{\xi}_k) - \hat{F}(z^k, \xi_k))\Big) \quad \text{with} \quad \bar{\xi}_k \sim \mathcal{P} \end{aligned} \tag{P_1SEG+}$$

The other variant applies the resolvent twice as in stochastic Mirror-Prox (Juditsky et al., 2011),

$$\begin{aligned} \bar{z}^k &= (\mathrm{id} + \gamma A)^{-1}(z^k - \gamma\hat{F}(z^k, \xi_k)) \quad \text{with} \quad \xi_k \sim \mathcal{P} \\ z^{k+1} &= (\mathrm{id} + \gamma A)^{-1}(z^k - \alpha_k\gamma\hat{F}(\bar{z}^k, \bar{\xi}_k)) \quad \text{with} \quad \bar{\xi}_k \sim \mathcal{P} \end{aligned} \tag{P_2SEG+}$$

When applying (EG+) and (SEG) to constrained settings we similarly use the following projected variants.

$$\begin{aligned} \bar{z}^k &= (\mathrm{id} + \gamma A)^{-1}(z^k - \gamma F z^k) \\ z^{k+1} &= (\mathrm{id} + \gamma A)^{-1}(z^k - \gamma F \bar{z}^k) \end{aligned} \tag{EG+}$$

and

$$\begin{aligned} \bar{z}^k &= (\mathrm{id} + \gamma A)^{-1}(z^k - \beta_k\gamma\hat{F}(z^k, \xi_k)) \quad \text{with} \quad \xi_k \sim \mathcal{P} \\ z^{k+1} &= (\mathrm{id} + \gamma A)^{-1}(z^k - \alpha_k\gamma\hat{F}(\bar{z}^k, \bar{\xi}_k)) \quad \text{with} \quad \bar{\xi}_k \sim \mathcal{P} \end{aligned} \tag{SEG}$$

H Comparison with variance reduction

Consider the case where the expectation comes in the form a finite sum,

$$Fz = \frac{1}{N}\sum_{\xi=1}^{N} \hat{F}(z,\xi). \tag{H.1}$$

In the worst case the averaged Lipschitz constant $F_{\hat{F}}$ scales proportionally to the number of elements N squared, i.e. $L_{\hat{F}} = \Omega(\sqrt{N}L_F)$. It is easy to construct such an example by taking one elements to have Lipschitz constant NL while letting the remaining elements have Lipschitz constant L. Recalling the definition in Assumption III, $L_{\hat{F}}^2 = \frac{N^2L^2}{N} + \frac{N-1}{N}L^2 \geq NL^2$ while the average becomes $L_F = \frac{NL}{N} + \frac{N-1}{N}L \leq 2L$ so $L_{\hat{F}} \geq \sqrt{N}/2L_F$. Thus, $L_{\hat{F}}$ can be $\sqrt{N}$ times larger than L_F, leading to a potentially strict requirement on the weak MVI parameter $\rho > -L_{\hat{F}}/2$ for variance reduction methods.