Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 18, No. 4, pp. 713–735, 2022

A VARIANCE-BASED PROXIMAL BACKWARD-FORWARD ALGORITHM WITH LINE SEARCH FOR STOCHASTIC MIXED VARIATIONAL INEQUALITIES*

Yue-Hong He and Xian-Jun Long†

Abstract: In this paper, we introduce a new variance-based proximal backward-forward algorithm with line search for stochastic mixed variational inequalities, which only needs to compute one proximal operator per iteration. Particularly, the proposed algorithm only requires the mapping F to be g-pseudomonotone and does not need to know any information of the Lipschitz constant of the mapping while other similar methods require the monotonicity and the information of the Lipschitz constant. Moreover, we analyse some properties of the proposed algorithm related to the asymptotic convergence, the linear convergence rate with finite computational budget and the optimal oracle complexity under some moderate conditions. Finally, some numerical experiments are given to show the efficiency and advantages of the algorithm introduced in this paper.

Key words: *stochastic mixed variational inequality, stochastic approximation, g-pseudomonotone, variance reduction, line search*

Mathematics Subject Classification: *65K15, 90C33, 90C15*

1 Introduction

Let $g : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper, convex and lower semicontinuous function. Denote by dom$g$ its domain, that is dom$g := \{x \in \mathbb{R}^n : g(x) < +\infty\}$. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping. The mixed variational inequality (MVI) problem is to find $x \in$ dom$g$ such that

$$\langle F(x), y - x\rangle + g(y) - g(x) \geq 0, \ \forall y \in \text{dom}g.$$

This problem is a generalization of variational inequalities and optimization problems and has wide applications in the fields of economics, structural engineering, physics and nonlinear programming. Numerous results have been obtained by many authors in the literatures; see [8, 9, 10, 11, 12, 20, 22, 23, 29, 30] and the references therein.

In fact, there exist a lot of problems that may be affected by uncertainties in real life, and ignoring these factors may cause serious economic or social damage. Besides, the MVI models with random variables are also used in many practical problems, for example, the

*This work was supported by the NSF of Chongqing (cstc2021jcyj-msxmX0721, cstc2018jcyjAX0119), the Education Committee Project Research Foundation of Chongqing (KJZDK201900801) and the Innovation Project for Graduate Students of Chongqing (CYS22629).

†Corresponding author.

game of suppliers and manufacturers where the price of raw materials provided by the suppliers is not fixed, and even may be affected by such as the weather, the market supply and demand, the mode of transportation and the international exchange rate. This implies that the existing algorithms for the MVI may not be applied directly to solve the problems with random factors. Therefore, it is of great significance to study the stochastic mixed variational inequalities under such realistic background and demand.

In this paper, we consider the following stochastic mixed variational inequality (SMVI) problem: find $x \in \mathrm{dom} g$ such that

$$\langle \mathbb{E}[f(x,\xi(\omega))], y-x\rangle + g(y) - g(x) \geq 0,\ \forall y \in \mathrm{dom} g, \tag{1.1}$$

where $\mathbb{E}$ denotes the expectation operator and $f(x,\xi) : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$ is measurable with respect to a random function $\xi : \Omega \to \Xi \subseteq \mathbb{R}^n$ defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and continuous with respect to x. The SMVI has received much attention due to its extensive applications in many fields such as stochastic optimization problems, stochastic saddle point problems and stochastic nash equilibrium problems; see [11-16] and the references therein. Denote by X^* the set of solutions for the SMVI. In what follows, we denote by $F(x) := \mathbb{E}[f(x,\xi(\omega))]$. If g is the indicator function of a closed and convex set X, then the SMVI reduces to the stochastic variational inequality (SVI) problem: find $x \in X$ such that

$$\langle \mathbb{E}[f(x,\xi(\omega))], y-x\rangle \geq 0,\ \forall y \in X.$$

Note that the probability distribution of the random variables may not be known completely in general, so the existing numerical methods for the SVI are not applicable to the SMVI directly. With this in mind, some methods such as the stochastic approximation (SA) method and the sample average approximation (SAA) method have been developed to deal with the involved expectation. Particularly, it has been shown that the SA method may be efficient in dealing with large-scale problems. To this end, the algorithm proposed in this paper for the SMVI is also based on the SA method.

The SA method was initially introduced by Robbins and Monro in [26] for solving stochastic root-finding problems. The first SA-based projection method for the SVI was introduced by Jiang and Xu [16] and they proved the convergence when the mapping F is strongly monotone and Lipschitz continuous. Yousefian et al. [34] proposed a SA-based projection scheme with distributed adaptive stepsize to deal with a Cartesian SVI in which the mapping F is still assumed to be strongly monotone and Lipschitz continuous. Koshal et al. [19] proposed a stochastic iterative Tikhonov regularization method for the SVI, its convergence requires the mapping F to be monotone and Lipschitz continuous. Iusem et al. [15] gave an incremental projection method with regularization for Cartesian SVI, whose convergence needs the mapping F to be monotone and Lipschitz continuous. Yousefian et al. [35] developed a regularized smoothing stochastic approximation scheme for the SVI where the mapping F is monotone and non-Lipschitzian. For more results on the SA method for the SVI, we refer to [7, 15, 17, 18, 31, 35, 36, 37] and the references therein.

Recently, to improve the iteration complexity and convergence speed, some improved SA-based methods were proposed by replacing the expected value function through the averages of increasing sample size of sampled functions. Particularly, Iusem et al. [13] developed a dynamic sampled SA-based extragradient method for the SVI, whose convergence requires the pseudo-monotonicity and Lipschitz continuity of the mapping. Since the Lipschitz constant may be unknown or it is difficult to estimate, Iusem et al. [14] proposed a dynamic sampled stochastic approximated extragradient method with line search for the SVI where the Lipschitz constant is not necessary to be known. Zhang et al. [42] proposed an infeasible single projection algorithm with line search for the pseudo-monotone SVI, which needs only one projection per iteration and the algorithm works without any information of the Lipschitz constant of the mapping. Recently, Yang et al. [38] developed a variance-based modified backward-forward algorithm with a stochastic approximate version of Armijo's line search for the pseudo-monotone SVI, which requires only one projection at each iteration. Based on this, one motivation of this paper is to extend the algorithm introduced in [38] to solve the SMVI under some moderate conditions.

Compared to the SVI, there are few papers to study the SMVI. In 2020, Mishchenko et al. [24] introduced an extended extragradient method to solve the SMVI and proved the convergence of their algorithm when the mapping F is monotone. Yang and Lin [32] proposed two fast variance-based proximal gradient algorithms for the SMVI, where both of them require two oracle calls per iteration. They investigated the convergence rate and the oracle complexity bound for the algorithms based on the monotonicity of the mapping F. Very recently, Yang and Lin [33] proposed three variance-based single-call proximal extragradient algorithms with high computational efficiency for the SMVI. Moreover, these algorithms require only one evaluation of the expected mapping at each iteration, but the convergence requires the mapping F to be monotone and Lipschitz continuous. It is obviously shown that the monotonicity assumption of the mapping F is a little strong, and many functions may not satisfy this condition in practice. Therefore, another motivation of this paper is to develop a SA-based algorithm to solve the SMVI problem under weaker assumptions.

Contributions. Inspired by the work reported in [13, 38, 32], to the best of our knowledge, we propose the first SA-based proximal backward-forward algorithm with respect to an unknown Lipschitz constant for solving the SMVI problem. Our contributions in this paper are five aspects:

(a) The proposed algorithm requires only one evaluation of the proximal mapping per iteration and can be seen as a significant extension of the work given by [13, 38] from the SVI to the SMVI. The convergence and the convergence rate are obtained without the assumption of pseudomonotonicity of the mapping F, while this assumption is essential in [13, 38].

(b) Compared with the algorithm introduced in [32] for the SMVI, we adopt a stochastic line search technique and do not require the information of the Lipschitz constant, which is beneficial to improve the computational efficiency.

(c) The operator involved is only g-pseudomonotone. This improves some recent results in [13, 38, 32] where the involved operator is monotone or pseudomonotone and therefore our algorithm is more applicable than the ones considered in [13, 38, 32] in practice.

(d) We analyse the asymptotic convergence, the optimal oracle complexity and the linear convergence rate with finite computational budget under the assumption of bounded proximal error bound.

(e) Some numerical experiments are given to illustrate the validity and superiority of the proposed algorithm in comparison with other algorithms in [13, 38, 32].

The rest of the paper is organized as follows: In Section 2, we give some related definitions and results. In Section 3, we describe the new algorithm and prove its asymptotic convergence. In Section 4, we discuss the linear convergence rate and oracle complexity of the proposed algorithm. Some numerical results are obtained in Section 5 to show the superiority of the algorithm and Section 6 obtains the conclusion.

2 Preliminaries

In this section, we recall some notations and results that may be used in the sequel. Let $\mathbb{R}^n$ be the n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$ and dom$g$ be a closed and convex set in $\mathbb{R}^n$. To ease notation, we use ξ to denote $\xi(\omega)$. For a σ-algebra $\mathcal{F}$, we denote by $\mathbb{E}[\xi]$ and $\mathbb{E}[\xi|\mathcal{F}]$ the expectation and conditional expectation, respectively. Besides, we denote by $|\xi|\mathcal{F}|_p := \sqrt[p]{\mathbb{E}[|\xi|_p|\mathcal{F}]}$ the $\mathcal{L}_p$-norm of ξ conditional to $\mathcal{F}$, where $|\xi|_p$ denotes the $\mathcal{L}_p$-norm of ξ with $p \geq 1$. Let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, where $\mathbb{N}$ is the set of all positive integers. For $b \in \mathbb{R}, \lceil b \rceil$ stands for the smallest integer no less than b. For a nonempty closed and convex set $C \subseteq \mathbb{R}^n, P_C$ denotes the projection operator onto C and σ_C denotes the indicator function of C; that is, $\sigma_C(x) = 0$ if $x \in C$ and ∞ otherwise. Let $g : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper lower semicontinuous convex function and a constant $\lambda > 0$, the proximal operator $\text{prox}_{\lambda g}$ is defined by

$$\text{prox}_{\lambda g}(x) = \underset{y \in \mathbb{R}^n}{\text{argmin}}\{g(y) + \frac{1}{2\lambda}\|x - y\|^2\},$$

where the domain of g is written as dom$g := \{x \in \mathbb{R}^n : g(x) < +\infty\}$. Clearly, $P_C(x) = \text{prox}_{\lambda \sigma_C}(x)$.

For a given number $p \geq 2$, we define the oracle error map $\varepsilon : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$ for (1.1) by

$$\varepsilon(x, \xi) := f(x, \xi) - F(x), \ \forall \xi \in \Xi, x \in \mathbb{R}^n. \tag{2.1}$$

and its p-moment function by

$$\delta_p(x) := \sqrt[p]{\mathbb{E}[\|\varepsilon(x, \xi)\|^p]}, \ \forall x \in \mathbb{R}^n.$$

Given an independent identically distributed (i.i.d) sample $\xi^N := \{\xi\}_{j=1}^N$ drawn from Ξ, we define two empirical mean operators and the oracle's empirical mean error associated to ξ^N, respectively, by

$$\widehat{F}(x, \xi^N) := \frac{1}{N}\sum_{j=1}^N f(x, \xi_j), \quad \widehat{F}(y, \eta^N) := \frac{1}{N}\sum_{j=1}^N f(y, \eta_j), \quad \widehat{\varepsilon}(x, \xi^N) := \frac{1}{N}\sum_{j=1}^N \varepsilon(x, \xi_j). \tag{2.2}$$

Assumption 2.1 ($A0$)**.**

(i) There exists a measurable function $\mathcal{L} : \Xi \to \mathbb{R}_+$ such that $\mathcal{L}(\xi) \geq 1$ for almost every $\xi \in \Xi$ and
$$\|f(x, \xi) - f(y, \xi)\| \leq \mathcal{L}(\xi)\|x - y\|, \ \forall x, y \in \mathbb{R}^n.$$

(ii) There exists $x^* \in \mathbb{R}^n$ and $p \geq 2$ such that $\mathbb{E}[\|f(x^*, \cdot)\|^p] < \infty$ and $\mathbb{E}[\mathcal{L}(\cdot)^p] < \infty$.

We now recall some lemmas which will be used later on.

Lemma 2.1 ([2])**.** *Let $g : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper, convex and lower semicontinuous function and $\lambda > 0$.*

(i) *For any $x, y \in \mathbb{R}^n$, $\|\text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y)\| \leq \|x - y\|$.*

(ii) *For all $y \in \mathbb{R}^n$, $p = \text{prox}_{\lambda g}(x)$ if and only if $(p - x)^T(y - p) \geq \lambda(g(p) - g(y))$. In particular, $(p - x)^T(y - p) \geq 0$ when $g(x) = \sigma_C(x), \forall y \in C$.*

(iii) *x^* is a solution of (1.1) if and only if* $x^* = \mathrm{prox}_{\lambda g}[x^* - \lambda F(x^*)]$.

Given $x \in \mathbb{R}^n$ and $\lambda > 0$, we define the natural residual function of (1.1) as

$$H_\lambda(x) := \left\| x - \mathrm{prox}_{\lambda g}[x - \lambda F(x)] \right\|. \tag{2.3}$$

Lemma 2.2 ([2])**.** *For all $x \in \mathbb{R}^n, \lambda \mapsto \frac{H_\lambda(x)}{\lambda}$ is non-increasing with respect to $\lambda \in (0, \infty)$.*

Lemma 2.3 ([14])**.** *Assume that (A0) holds. For all $q \in [p, 2p]$, the mean operator F is L-Lipschitz continuous on $\mathbb{R}^n$ and δ_q is L_q-Lipschitz continuous on $\mathbb{R}^n$, where $L := \mathbb{E}[\mathcal{L}(\cdot)]$ and $L_q := \sqrt[q]{\mathbb{E}[\mathcal{L}(\cdot)^q]} + L$.*

Lemma 2.4 ([4])**.** *For any $q \geq 2$ and vector-valued martingale $\{y_j\}_{j=0}^N$ adapted to the filtration $\{\mathcal{F}_j\}_{j=0}^N$ with $y_0 = 0$, there exists $C_q > 0$ such that*

$$\left| \sup_{j \leq N} \|y_j\| \right|_q \leq C_q \left| \sqrt{\sum_{j=1}^N \|y_j - y_{j-1}\|^2} \right|_q \leq C_q \sqrt{\sum_{j=1}^N | \, \|y_j - y_{j-1}\| \, |_q^2}.$$

Lemma 2.5. *Let $\xi^N := \{\xi_j\}_{j=1}^N$ be an i.i.d. sample drawn from Ξ. Assume that (A0) holds and take $q \in [p, 2p]$. Set $C_2 := 1$ if $q = p = 2$. Otherwise, C_q is defined as in Lemma 2.4. Then, for any $x \in \mathbb{R}^n$ and $x^* \in \mathrm{dom} g$, we have*

$$| \, \|\widehat{\varepsilon}(x, \xi^N)\| \, |_q \leq C_q \frac{\delta_q(x^*) + L_q \|x - x^*\|}{\sqrt{N}}.$$

Proof. Since dom$g$ is a closed convex set, we can replace X in Lemma 3.9 of [14] by dom$g$, hence the proof is similar to the one in [14]. □

Lemma 2.6. *Let $\xi^N := \{\xi_j\}_{j=1}^N$ is an i.i.d. sample drawn from Ξ and let $\alpha_N : \Xi \to [0, \gamma]$ be a random variable for some $0 < \gamma \leq 1$. Given $(\alpha, x) \in [0, \gamma] \times \mathbb{R}^n$ and $y(x, \alpha, \xi^N) := \mathrm{prox}_{\alpha g}[x - \alpha \widehat{F}(x, \xi^N)]$. Assume that (A0) holds. Then there exist positive constants $\{c_i\}_{i=1}^4$ such that, for any $x \in \mathbb{R}^n$ and $x^* \in X^*$,*

$$| \, \|\widehat{\varepsilon}(y(x, \alpha_N, \xi^N), \xi^N)\| \, |_p \leq \frac{c_1 \delta_{2p}(x^*) + \widehat{L}_{2p} \|x - x^*\|}{\sqrt{N}},$$

where $\widehat{L}_{2p} := c_2 L_2 + c_3 L_p + c_4 L_{2p}$.

Proof. Note that dom$g$ is a closed and convex set. Let $x \in \mathrm{dom} g$ and $x^* \in X^*$. Set $y^N := y(x, \alpha_N, \xi^N)$. For any $s > 0$, let $R(s) := (1 + L\gamma)\|x - x^*\| + \gamma s$, and denote by $\mathbb{B}(s) := \mathbb{B}[x^*, R(s)]$ the ball.

Example 14.29 of [27] and (A0) imply that the map $\Xi \times \mathrm{dom} g \ni (\omega, x) \longmapsto \|\widehat{\varepsilon}(x, \xi^N(\omega))\|$ is a normal integrand, that is,

$$\omega \longmapsto \mathrm{epi}\|\widehat{\varepsilon}(x', \xi^N(\omega))\| := \{(x, y) \in \mathrm{dom} g \times \mathbb{R} : \|\widehat{\varepsilon}(x, \xi^N(\omega))\| \leq y\}$$

is a set-valued measurable function $\varepsilon : \Omega \longrightarrow [0, \infty)$ and $R > 0$. It follows that

$$\omega \longmapsto \sup_{x' \in \mathbb{B}(\varepsilon(\omega)) \cap domg} \|\widehat{\varepsilon}(x', \xi^N(\omega))\| \text{ and } \omega \longmapsto \sup_{x' \in \mathbb{B}[x^*, R] \cap domg} \|\widehat{\varepsilon}(x', \xi^N(\omega))\|$$

are measurable functions.

When dom$g$ is unbounded, given $\alpha \in [0,\gamma]$, Lemma 2.1 (iii) implies that $x^* = \text{prox}_{\alpha g}[x^* - \alpha F(x^*)]$. Then, from Lemma 2.1 (i), the definitions of $y(x,\alpha,\xi^N)$, (2.1) and (2.2), we get that, for any $\alpha \in [0,\gamma]$,

$$\begin{aligned}\|x^* - y(x,\alpha,\xi^N)\| =& \|\text{prox}_{\alpha g}[x^* - \alpha F(x^*)] - \text{prox}_{\alpha g}[x - \alpha(F(x) + \widehat{\varepsilon}(x,\xi^N))]\| \\ \le& \|x - x^*\| + \alpha\|F(x) - F(x^*)\| + \alpha\|\widehat{\varepsilon}(x,\xi^N)\| \\ \le& (1 + L\gamma)\|x - x^*\| + \gamma\|\widehat{\varepsilon}(x,\xi^N)\|,\end{aligned}$$

where, in the last inequality, we used the Lipschitz continuity of F. The rest of the proof is similar to the proof of Theorem 3.11 in [14]. □

Remark 2.7. In Lemma 2.6, the constants satisfy

$$c_1 := 2C_p + C_{2p}C_{\gamma L,p}, \quad c_2 \lesssim \left(\frac{3\sqrt{n}}{\sqrt{2}-1} + \sqrt{p}\right) C_{\gamma L,p}, \quad c_3 \lesssim pC_{\gamma L,p}, \quad c_4 := C_{2p}C_{\gamma L,p},$$

where $C_{\gamma L,p} := 1 + 2\gamma L + \gamma|\mathcal{L}(\xi)|_{2p}$ and $\{C_p, C_{2p}\}$ are defined as in Lemma 2.6.

Lemma 2.8 ([25])**.** *Let $\{v^k\},\{u^k\},\{a^k\},\{b^k\}$ be sequences of non-negative random variables, adapted to the filtration $\{\mathcal{F}_k\}$ such that almost surely $\sum_{k=1}^{\infty} a^k < \infty, \sum_{k=1}^{\infty} b^k < \infty$ and for each $k \in \mathbb{N}, \mathbb{E}[v^{k+1}|\mathcal{F}_k] \le (1 + a^k)v^k - u^k + b^k$. Then almost surely $\{v^k\}$ converges and $\sum_{k=1}^{\infty} u^k < \infty$.*

3 Algorithm and Convergence Analysis

In this section, we propose the variance-based proximal backward-forward algorithm with line search for solving the SMVI and then prove the convergence of the proposed algorithm.

3.1 Algorithm

In this subsection, motivated by the works reported in [13, 38, 32], we introduce the following variance-based proximal backward-forward algorithm with line search for solving the SMVI.

Remark 3.1. (a) Compared with the algorithm proposed in [13], Algorithm 1 requires only one proximal projection at each iteration, which is beneficial to reduce the computational cost.

(b) In Algorithm 1, we use the dynamic sampled SA line search scheme to ensure the step size so that the termination criteria is generally not clear. The solution to this problem is that Algorithm 1 regenerates a sample in Step 1 if $x^k = \text{prox}_{\gamma g}[x^k - \gamma\widehat{F}(x^k,\xi^k)]$, which makes the line search step (3.1) terminated after a finite number of steps. Note that the algorithm terminates if the maximum iteration number is reached when the algorithm is implemented in practice.

(c) For deterministic variational inequalities, x^k is an exact solution when $x^k = \Pi_X[x^k - \gamma F(x^k)]$, so there is no need to regenerate a random sample in Step 1.

Proposition 3.2. *Assume that (A0) holds. The line search step (3.1) in the iteration k of Algorithm 1 terminates after a finite number l_k of steps.*

Algorithm 1 (Variance-based proximal backward-forward algorithm with line search)

Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$, parameters $\beta \in (1-\frac{1}{\sqrt{2}}, 1+\frac{1}{\sqrt{2}}), \mu \in (0, \sqrt{\frac{2\beta-\beta^2-\frac{1}{2}}{3\beta^2}}), \gamma, \theta \in (0,1)$, and the sample rate $\{N_k\} \subset \mathbb{N}$. Set $k := 0$.

Step 1: Generate a sample $\xi^k := \{\xi_j^k\}_{j=1}^{N_k}$ from Ξ and compute $\widehat{F}(x^k, \xi^k) := \frac{1}{N_k}\sum_{j=1}^{N_k} f(x^k, \xi_j^k)$. If $x^k = \operatorname{prox}_{\gamma g}[x^k - \gamma \widehat{F}(x^k, \xi^k)]$, regenerate a sample. Otherwise, go to **Step 2**.

Step 2. Choose α_k as the maximum $\alpha \in \{\gamma\theta^{l_k} | l_k \in \mathbb{N}_0\}$ such that

$$\alpha \left\| \widehat{F}(y^k(\alpha), \xi^k) - \widehat{F}(x^k, \xi^k) \right\| \le \mu \left\| y^k(\alpha) - x^k \right\|, \tag{3.1}$$

where $y^k(\alpha) := \operatorname{prox}_{\alpha g}[x^k - \alpha \widehat{F}(x^k, \xi^k)]$ and $\widehat{F}(y^k(\alpha), \xi^k) := \frac{1}{N_k}\sum_{j=1}^{N_k} f(y^k(\alpha), \xi_j^k)$.

Step 3. Set $y^k := \operatorname{prox}_{\alpha_k g}[x^k - \alpha_k \widehat{F}(x^k, \xi^k)]$ and generate a sample $\eta^k := \{\eta_j^k\}_{j=1}^{N_k}$ from Ξ. Calculate the next iterate

$$x^{k+1} := (1-\beta)x^k + \beta(y^k + \alpha_k(\widehat{F}(x^k, \xi^k) - \widehat{F}(y^k, \eta^k))). \tag{3.2}$$

Set $k := k+1$ and return to **Step 1**.

Proof. We consider the following two situations. (i) If $x^k \in \operatorname{dom} g$, the conclusion follows from Lemma 3.6 in [14] immediately with $d^k = 0$. (ii) If $x^k \notin \operatorname{dom} g$, we set $\alpha_l := \gamma\theta^{l_k}$ and $F(\cdot) := \frac{1}{N_k} f(\cdot, \xi_j^k)$ in (2.3). Suppose by contradiction that the line search (3.1) dose not terminate in a finite number of iterations. That is, for every $l_k \in \mathbb{N}_0$, we have

$$\alpha_l \left\| \widehat{F}(y^k(\alpha_l), \xi^k) - \widehat{F}(x^k, \xi^k) \right\| > \mu H_{\alpha_l}(x_k).$$

Taking the limit $l_k \to \infty$ in the above inequality, we have from the continuity of $\widehat{F}(\cdot, \xi^k)$ and $\operatorname{prox}_{\lambda g}(\cdot)$ that $0 \ge \mu \|x^k - \operatorname{prox}_{\alpha_l, g}[x^k]\| > 0$. This gives a contradiction. Therefore, the line search step (3.1) terminates in a finite number of iterations. □

3.2 Almost sure convergence

In this subsection, we will prove the almost sure convergence of Algorithm 1. To this end, we first define the filtrations

$$\mathcal{F}_k := \sigma(x^0, \xi^0, \xi^1, \cdots, \xi^{k-1}, \eta^0, \cdots, \eta^{k-1}), \quad \widehat{\mathcal{F}}_k = \sigma(x^0, \xi^0, \xi^1, \cdots, \xi^k, \eta^0, \cdots, \eta^{k-1})$$

and the oracle errors

$$\hat{\varepsilon}_1^k := \widehat{\varepsilon}(x^k, \xi^k), \qquad \hat{\varepsilon}_2^k := \widehat{\varepsilon}(y^k, \eta^k), \qquad \hat{\varepsilon}_3^k := \widehat{\varepsilon}(y^k, \xi^k).$$

We need the following assumptions which will be used later on.

Assumption 3.1 ($A1$)**.** The solution set X^* of (1.1) is nonempty.

Assumption 3.2 ($A2$)**.** The mean operator $F : \operatorname{dom} g \to \mathbb{R}^n$ is g-pseudomonotone on $\operatorname{dom} g$, that is

$$\langle F(x), y-x \rangle + g(y) - g(x) \ge 0 \Longrightarrow \langle F(y), y-x \rangle + g(y) - g(x) \ge 0, \ \forall x, y \in \operatorname{dom} g.$$

Remark 3.3. Clearly, a monotone mapping is pseudomonotone and g-pseudomonotone, but the converse is not true in general, see example 3.1. Moreover, if g=0, then a g-pseudomonotone mapping reduces to a pseudomonotone mapping. Here, we give the following two examples satisfying the g-pseudomonotonicity.

The first example is a variant of the Example 2.3 of [43], which shows that a mapping possesses g-pseudomonotonicity property, but is not monotone in the usual sense.

Example 3.1 Let dom$g = (-\infty, +\infty)$ and $D = [a, b] \subset \mathbb{R}$ with $0 \leq a < c \leq b$ be a closed and convex subset. Let $g : D \longrightarrow$ dom$g$ be a differentiable and convex function such that, $|g'(x)| \leq r$ for all $x \in D$, where $r > 0$ is a constant. Let

$$F(x) = \begin{cases} r + 5 + e^x, \ x \in [a, c), \\ \\ r + 2 + \sin x, \ x \in [c, b], \end{cases}$$

It is easy to check that F is not monotone on D. However, we can show that F is g-pseudomonotone on D. In fact, let

$$\langle F(x), y - x\rangle + g(y) - g(x) \geq 0.$$

Then by the Mean Value Theorem, there exists $\xi \in [a, b]$ such that, $g(y)-g(x) = g'(\xi)(y-x)$ and so

$$\langle F(x), y - x\rangle + g'(\xi)(y - x) \geq 0.$$

Thus,

$$(F(x) + g'(\xi))(y - x) \geq 0.$$

We know that $F(x) \geq r + 1$ and so $F(x) + g'(\xi) > 0$. It follows from the above inequality that $y - x \geq 0$. Hence, we have

$$\langle F(y), y - x\rangle + g(y) - g(x) = (F(y) + g'(\xi))(y - x) \geq 0,$$

which shows that F is g-pseudomonotone on D.

The second example is a variant of the Example 2.1 of [43], which shows that a g-pseudomonotone mapping is not necessary pseudomonotone if $g \neq 0$.

Example 3.2 Let dom$g = (-\infty, +\infty)$ and $D = [2, 4]$. Let

$$g(x) = x^2, \ x \in D$$

and

$$F(x) = \begin{cases} \frac{1}{2}x, \ x \in [2, 3], \\ \\ \frac{1}{2}x - 3, \ x \in (3, 4]. \end{cases}$$

It is easy to see that F is not pseudomonotone on D. However, we can show that F is g-pseudomonotone on D. In fact, let

$$\langle F(x), y - x\rangle + g(y) - g(x) = (F(x) + y + x)(y - x) \geq 0.$$

It follows that $F(x) + y + x > 0$ and so $y - x \geq 0$. Thus, we have

$$\langle F(y), y - x\rangle + g(y) - g(x) = (F(y) + y + x)(y - x) \geq 0.$$

This implies that F is g-pseudomonotone on D.

Assumption 3.3 ($A3$)**.** In Algorithm 1, the sequence $\{\xi_j^k | k \in \mathbb{N}_0, j \in [N_k]\}$ and $\{\eta_j^k | k \in \mathbb{N}_0, j \in [N_k]\}$ are i.i.d. samples drawn from Ξ independent of each other. Moreover, $\sum_{k=0}^{\infty} \frac{1}{N_k} < \infty$.

Remark 3.4. In this paper, we set $N_k = \mathcal{N}\lceil (k+\lambda)(\ln(k+\lambda))^{1+b}\rceil$ and this can be regarded as a sufficient choice of the above assumption, where $\mathcal{N} \in \mathbb{N}, \lambda > 0$ and $b > 0$.

Lemma 3.5 ([38])**.** *Let* $\widehat{L}_k := \frac{1}{N_k}\sum_{j=1}^{N_k} \mathcal{L}(\xi_j^k)$. *If* ($A0$) *and* ($A3$) *hold, then almost surely* $\alpha_k \geq \min\{\frac{\mu\theta}{\widehat{L}_k}, \gamma\}$, *and* $|\alpha_k|\mathcal{F}_k|_2 \cdot |\mathcal{L}(\xi)|_2 \geq \min\{\mu\theta, \gamma\}$.

Remark 3.6. By Lemma 3.5, the step size sequence $\{\alpha_k\}$ generated by Algorithm 1 satisfies $|\alpha_k|\mathcal{F}_k|_2 \cdot |\mathcal{L}(\xi)|_2 \geq \min\{\mu\theta, \gamma\}$ and $\alpha_k < 1 (\forall k \in \mathbb{N}_0)$; that is, the conditional expectation to $\mathcal{F}_k$ of the sequence $\{\alpha_k\}$ is bounded.

The following lemma is essential to the analysis of convergence of the proposed algorithm.

Lemma 3.7 (a recursive error bound)**.** *Assume that* ($A0$) *and* ($A2$) *hold. Let* $\{x_k\}$ *and* $\{y_k\}$ *be the sequences generated by Algorithm* 1. *Then, for any* $x^* \in X^*$,

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2 \leq \left\|x^k - x^*\right\|^2 - \frac{A_k}{2}\alpha_k^2 H(x^k)^2 + A_k\gamma^2\left\|\hat{\varepsilon}_1^k\right\|^2 \\ + 6\beta^2\gamma^2\left\|\hat{\varepsilon}_2^k\right\|^2 + 6\beta^2\gamma^2\left\|\hat{\varepsilon}_3^k\right\|^2 + 2\beta\alpha_k\langle x^* - y^k, \hat{\varepsilon}_2^k\rangle,\end{aligned} \tag{3.3}$$

where $A_k = 2\beta(2 - \beta - 3\beta\mu^2) - 1$.

Proof. For any $x^* \in X^*$ and $k \in \mathbb{N}_0$, we have

$$\begin{aligned}\left\|x^k - x^*\right\|^2 &= \left\|x^k - y^k + y^k - x^{k+1} + x^{k+1} - x^*\right\|^2 \\ &= \left\|x^k - y^k\right\|^2 + \left\|y^k - x^{k+1}\right\|^2 + \left\|x^{k+1} - x^*\right\|^2 + 2\langle x^k - y^k, y^k - x^{k+1}\rangle \\ &\quad + 2\langle x^k - y^k, x^{k+1} - x^*\rangle + 2\langle y^k - x^{k+1}, x^{k+1} - x^*\rangle \\ &= \left\|x^k - y^k\right\|^2 + \left\|y^k - x^{k+1}\right\|^2 + \left\|x^{k+1} - x^*\right\|^2 + 2\langle x^k - y^k, y^k - x^*\rangle \\ &\quad + 2\langle y^k - x^{k+1}, x^{k+1} - x^*\rangle \\ &= \left\|x^k - y^k\right\|^2 + \left\|y^k - x^{k+1}\right\|^2 + \left\|x^{k+1} - x^*\right\|^2 + 2\langle x^k - y^k, y^k - x^*\rangle \\ &\quad - 2\left\|y^k - x^{k+1}\right\|^2 + 2\langle y^k - x^{k+1}, y^k - x^*\rangle \\ &= \left\|x^k - y^k\right\|^2 - \left\|y^k - x^{k+1}\right\|^2 + \left\|x^{k+1} - x^*\right\|^2 + 2\langle x^k - x^{k+1}, y^k - x^*\rangle.\end{aligned}$$

This implies that

$$\left\|x^{k+1} - x^*\right\|^2 = \left\|x^k - x^*\right\|^2 - \left\|x^k - y^k\right\|^2 + \left\|y^k - x^{k+1}\right\|^2 - 2\langle x^k - x^{k+1}, y^k - x^*\rangle.$$

By the definition of y^k and Lemma 2.1,

$$\langle y^k - x^k + \alpha_k(F(x^k) + \hat{\varepsilon}_1^k), x^* - y^k\rangle \geq \alpha_k(g(y^k) - g(x^*)).$$

It follows that

$$\begin{aligned}\left\|x^{k+1}-x^*\right\|^2 &= \left\|x^k-x^*\right\|^2 - \left\|x^k-y^k\right\|^2 \\ &\quad + \left\|(1-\beta)(y^k-x^k) - \beta\alpha_k(\widehat{F}(x^k,\xi^k) - \widehat{F}(y^k,\eta^k))\right\|^2 \\ &\quad - 2\beta\langle x^k-y^k-\alpha_k(F(x^k)+\hat{\varepsilon}_1^k - F(y^k) - \hat{\varepsilon}_2^k), y^k-x^*\rangle \\ &\le \left\|x^k-x^*\right\|^2 - \left\|x^k-y^k\right\|^2 + 2(1-\beta)^2\left\|x^k-y^k\right\|^2 \\ &\quad + 2\beta^2\alpha_k^2\left\|\widehat{F}(x^k,\xi^k)-\widehat{F}(y^k,\eta^k)\right\|^2 \\ &\quad - 2\beta\alpha_k(\langle F(y^k), y^k-x^*\rangle + g(y^k)-g(x^*)) - 2\beta\alpha_k\langle\hat{\varepsilon}_2^k, y^k-x^*\rangle \\ &\le \left\|x^k-x^*\right\|^2 + (2(1-\beta)^2-1)\left\|x^k-y^k\right\|^2 \\ &\quad + 6\beta^2\alpha_k^2\left\|\widehat{F}(x^k,\xi^k)-\widehat{F}(y^k,\xi^k)\right\|^2 \\ &\quad + 6\beta^2\alpha_k^2\left\|\widehat{F}(y^k,\xi^k)-F(y^k)\right\|^2 + 6\beta^2\alpha_k^2\left\|\widehat{F}(y^k,\eta^k)-F(y^k)\right\|^2 \\ &\quad - 2\beta\alpha_k(\langle F(y^k), y^k-x^*\rangle + g(y^k)-g(x^*)) - 2\beta\alpha_k\langle\hat{\varepsilon}_2^k, y^k-x^*\rangle \\ &\le \left\|x^k-x^*\right\|^2 - A_k\left\|x^k-y^k\right\|^2 + 6\beta^2\gamma^2\left\|\hat{\varepsilon}_2^k\right\|^2 + 6\beta^2\gamma^2\left\|\hat{\varepsilon}_3^k\right\|^2 \\ &\quad - 2\beta\alpha_k(\langle F(y^k), y^k-x^*\rangle + g(y^k)-g(x^*)) - 2\beta\alpha_k\langle\hat{\varepsilon}_2^k, y^k-x^*\rangle.\end{aligned} \tag{3.4}$$

On the other hand, we have from $x^* \in X^*$ and $(A2)$ that

$$\langle F(y^k), y^k-x^*\rangle + g(y^k) - g(x^*) \ge 0. \tag{3.5}$$

Combining (3.4) and (3.5) yields

$$\begin{aligned}\left\|x^{k+1}-x^*\right\|^2 &\le \left\|x^k-x^*\right\|^2 - A_k\left\|x^k-y^k\right\|^2 \\ &\quad + 6\beta^2\gamma^2\left\|\hat{\varepsilon}_2^k\right\|^2 + 6\beta^2\gamma^2\left\|\hat{\varepsilon}_3^k\right\|^2 + 2\beta\alpha_k\langle x^*-y^k, \hat{\varepsilon}_2^k\rangle.\end{aligned} \tag{3.6}$$

Note that $y^k = \mathrm{prox}_{\alpha_k g}[x^k - \alpha_k(F(x^k)+\hat{\varepsilon}_1^k)]$. By Lemmas 2.1 and 2.2,

$$\begin{aligned}\alpha_k^2 H(x^k)^2 &\le H_{\alpha_k}(x^k)^2 \\ &= \left\|x^k - \mathrm{prox}_{\alpha_k g}[x^k-\alpha_k F(x^k)]\right\|^2 \\ &\le 2\left\|x^k-y^k\right\|^2 + 2\left\|\mathrm{prox}_{\alpha_k g}[x^k-\alpha_k(F(x^k)+\hat{\varepsilon}_1^k)] - \mathrm{prox}_{\alpha_k g}[x^k-\alpha_k F(x^k)]\right\|^2 \\ &\le 2\left\|x^k-y^k\right\|^2 + 2\gamma^2\left\|\hat{\varepsilon}_1^k\right\|^2.\end{aligned}$$

This together with (3.6) yields the conclusion. □

Remark 3.8. In Lemma 3.2, we obtain a desired recursive relation with three oracle calls per iteration, which only requires the g-pseudomonotonicity of the mapping, while the algorithm considered in [13, 38, 32] requires that the mapping is monotone or pseudomonotone. Besides, the variation interval of μ with respect to the line search scheme is larger than that in [38]. Therefore, the above advantages of the proposed algorithm indicate that our algorithm is more applicable and effective than other algorithms considered in [13, 38, 32] in practice.

We next aim at controlling the bound of the term $A_k\gamma^2\left\|\hat{\varepsilon}_1^k\right\|^2+6\beta^2\gamma^2\left\|\hat{\varepsilon}_2^k\right\|^2+6\beta^2\gamma^2\left\|\hat{\varepsilon}_3^k\right\|^2$ in the right side of (3.3). Since $\hat{\varepsilon}_1^k$ and $\hat{\varepsilon}_2^k$ are martingale differences, we can control them in a relatively straightforward way by Lemma 2.5. The non-martingale difference $\hat{\varepsilon}_3^k$ is controlled by Lemma 2.6 based on the empirical process theory that is mentioned in [14]. Then, we present some lemmas to show the convergence of the proposed algorithm.

Lemma 3.9 (bound on oracle error)**.** *Assume that (A0), (A1) and (A3) hold. Then there exist positive constants C_p and $\bar{C}_p$ (depending only on $\{n,p,\gamma\mathcal{L}(\xi),\sup_k\frac{1}{N_k}\}$) such that, for all $x^*\in X^*$,*

$$\triangle B_k\le\frac{C_p\gamma^2\delta_{2p}^2(x^*)+\bar{C}_p^2\gamma^2\widehat{L}_{2p}^2\left\|x^k-x^*\right\|^2}{N_k},$$

where $\triangle B_k:=A_k\gamma^2\big|\left\|\hat{\varepsilon}_1^k\right\|^2|\mathcal{F}_k\big|_{\frac{p}{2}}+6\beta^2\gamma^2\big|\left\|\hat{\varepsilon}_2^k\right\|^2|\mathcal{F}_k\big|_{\frac{p}{2}}+6\beta^2\gamma^2\big|\left\|\hat{\varepsilon}_3^k\right\|^2|\mathcal{F}_k\big|_{\frac{p}{2}}$, $C_p:=2c_1^2[A_k+12\beta^2+\sup_k\frac{24\gamma^2\beta^2\widehat{L}_{2p}^2}{N_k}]$, $\bar{C}_p:=2(A_k+6\beta^2+12\beta^2(1+\gamma L)^2+\sup_k\frac{24\gamma^2\beta^2\widehat{L}_{2p}^2}{N_k})$.

Proof. Since ξ^k is independent of $\mathcal{F}_k$, we have from $x^k\in\mathcal{F}_k$ and Lemma 2.5 with $p=q$ that

$$\big|\left\|\hat{\varepsilon}_1^k\right\||\mathcal{F}_k\big|_p\le C_p\frac{\delta_p(x^*)+L_p\left\|x^k-x^*\right\|}{\sqrt{N_k}}. \tag{3.7}$$

Note that

$$\begin{aligned}\left\|x^*-y^k\right\|&=\left\|\mathrm{prox}_{\alpha_k g}[x^*-\alpha_kF(x^*)]-\mathrm{prox}_{\alpha_k g}[x^k-\alpha_k(F(x^k)+\hat{\varepsilon}_1^k)]\right\|\\&\le\left\|x^*-x^k-\alpha_k(F(x^*)-F(x^k))+\alpha_k\hat{\varepsilon}_1^k\right\|\\&\le\left\|x^*-x^k\right\|+\alpha_k\left\|F(x^*)-F(x^k)\right\|+\alpha_k\left\|\hat{\varepsilon}_1^k\right\|\\&\le(1+\gamma L)\left\|x^*-x^k\right\|+\gamma\left\|\hat{\varepsilon}_1^k\right\|.\end{aligned}$$

Taking $\big|\cdot|\mathcal{F}_k\big|_p$ in the above inequality, we obtain

$$\big|\left\|x^*-y^k\right\||\mathcal{F}_k\big|_p\le(1+\gamma L)\left\|x^*-x^k\right\|+\gamma\big|\left\|\hat{\varepsilon}_1^k\right\||\mathcal{F}_k\big|_p. \tag{3.8}$$

Note that η^k is independent of $\mathcal{F}_k$ and $||\cdot|\widehat{\mathcal{F}}_k|_p|\mathcal{F}_k|_p=\big|\cdot|\mathcal{F}_k\big|_p$, it follows from $y^k\in\mathcal{F}_k$ and Lemma 2.5 with $p=q$ that

$$|\left\|\hat{\varepsilon}_2^k\right\||\mathcal{F}_k|_p=\big|\big|\left\|\hat{\varepsilon}_2^k\right\||\widehat{\mathcal{F}}_k\big|_p|\mathcal{F}_k\big|_p\le C_p\frac{\delta_p(x^*)+L_p\big|\left\|y^k-x^*\right\||\mathcal{F}_k\big|_p}{\sqrt{N_k}}. \tag{3.9}$$

Moreover, from Lemma 2.6, $0<\alpha_k\le\gamma<1$, $y^k=y(x^k,\alpha_k,\xi^k)$, $x^k\in\mathcal{F}_k$ and ξ^k is independent of $\widehat{\mathcal{F}}_k$, we have

$$\big|\left\|\hat{\varepsilon}_3^k\right\||\mathcal{F}_k\big|_p\le\frac{c_1\delta_{2p}(x^*)+\widehat{L}_{2p}\left\|x^k-x^*\right\|}{\sqrt{N_k}}. \tag{3.10}$$

Since $|a^2|\mathcal{F}_k|_{\frac{p}{2}}=|a|\mathcal{F}_k|_p^2$, $(a+b)^2\le 2a^2+2b^2$, $\widehat{L}_{2p}>L_pC_p$, $c_1>C_p$ and $\delta_{2p}(x^*)\ge\delta_p(x^*)$, by (3.7)-(3.10) we can obtain the conclusion of Lemma 3.9. ∎

Lemma 3.10 (stochastic quasi-Fejér property)**.** *Assume that* $(A0)-(A3)$ *hold. Let* $p=2$ *in Lemma 3.9 and* $\rho := \frac{A_k(min\{\mu\theta,\gamma\})^2}{2|L(\xi)|_2^2}$. *Then, for all* $x^* \in X^*$ *and* $k \in \mathbb{N}_0$, *we have*

$$\mathbb{E}[\left\|x^{k+1}-x^*\right\|^2|\mathcal{F}_k] \leq \left\|x^k-x^*\right\|^2 - \rho H(x^k)^2 + \frac{C_2\gamma^2\delta_4^2(x^*) + \bar{C}_2\gamma^2\widehat{L}_4^2\left\|x^k-x^*\right\|^2}{N_k}, \quad (3.11)$$

where C_2 *and* $\bar{C}_2$ *are the same as in Lemma* 3.9.

Proof. Since $y^k \in \widehat{\mathcal{F}}_k$ and η^k is independent of $\widehat{\mathcal{F}}_k$,

$$\mathbb{E}[\hat{\varepsilon}_2^k|\widehat{\mathcal{F}}_k] = \frac{1}{N_k}\sum_{j=1}^{N_k}\mathbb{E}[f(y^k,\eta_j^k)|\widehat{\mathcal{F}}_k] - F(y^k) = \frac{1}{N_k}\sum_{j=1}^{N_k}F(y^k) - F(y^k) = 0.$$

From $y^k \in \widehat{\mathcal{F}}_k$ and $a_k \in \widehat{\mathcal{F}}_k$, we obtain $\mathbb{E}[a_k(x^*-y^k)^T\hat{\varepsilon}_2^k|\widehat{\mathcal{F}}_k]=0$. Note that $\mathbb{E}[\mathbb{E}[\cdot|\widehat{\mathcal{F}}_k]|\mathcal{F}_k] = \mathbb{E}[\cdot|\mathcal{F}_k]$. It follows that $\mathbb{E}[a_k\langle x^*-y^k,\hat{\varepsilon}_2^k\rangle|\mathcal{F}_k]=0$. Taking $\mathbb{E}[\cdot|\mathcal{F}_k]$ in (3.3), we obtain the conclusion from the facts of $x^k \in \mathcal{F}_k$ and Lemmas 3.5, 3.7 and 3.9. □

Theorem 3.11 (asymptotic convergence)**.** *Assume that (A0)-(A3) hold. Then, almost surely the sequence* $\{x^k\}$ *generated by Algorithm 1 is bounded,* $\lim_{k\to\infty} dist(x^k,X^*)=0$ *and* $H(x^k)\to 0$. *In particular, almost surely every cluster point of* $\{x^k\}$ *belongs to* X^*.

Proof. Let $x^* \in X^*$. Note that $\sum_{k=0}^{\infty}\frac{1}{N_k}<\infty$ and $x^k \in \mathcal{F}_k$. By applying Lemma 2.8 with $v^k := \left\|x^k-x^*\right\|^2, a_k := \frac{\bar{C}_2\gamma^2\widehat{L}_4^2}{N_k}, b_k := \frac{C_2\gamma^2\delta_4^2(x^*)}{N_k}$ and $u_k := \rho H(x^k)^2$, we have that, almost surely, $\{\left\|x^k-x^*\right\|^2\}$ converges and $\sum_{k=0}^{\infty}H(x^k)^2<\infty$. In particular, almost surely $\{x^k\}$ is bounded and

$$0 = \lim_{k\to\infty}H(x^k)^2 = \lim_{k\to\infty}\left\|x^k - \text{prox}_g[x^k-F(x^k)]\right\|^2. \quad (3.12)$$

From the continuity of F and the proximal mapping, we know that almost surely every cluster point $\bar{x}$ of $\{x^k\}$ satisfies $\left\|\bar{x}-\text{prox}_g[\bar{x}-F(\bar{x})]\right\|^2=0$. Furthermore, by Lemma 2.1, we have $\bar{x}\in X^*$. It follows that the boundedness of $\{x^k\}$ and the fact that every cluster point of $\{x^k\}$ belonging to X^* yield that $\lim_{k\to\infty} dist(x^k,X^*)=0$. Similarly, we can deduce that $\lim_{k\to\infty}\mathbb{E}[H(x^k)^2]=0$ by taking expectation in (3.11). □

Remark 3.12. In [38], Yang et al. also obtained the asymptotic convergence of their algorithm. The related properties in [38] require the pseudomonotonicity of the mapping, which is more restrictive than the assumption of g-pseudomonotonicity in Algprithm 1. Moreover, those results obtained in [38] are only used to solve the SVI, while our results can be seen as a significant extension of the work given by [13, 38] from the SVI to the SMVI.

4 Convergence Rate and Oracle Complexity Analysis

In this section, we analyse the optimal oracle complexity and the sublinear convergence rate in terms of the mean natural residual function. Moreover, we discuss the linear convergence rate for the proposed algorithm under the bounded proximal error bound condition.

We first prove the following lemma.

Lemma 4.1 ($\mathcal{L}^2$-boundedness of the iterates)**.** *Assume that* $(A0)-(A3)$ *hold and* $\{x^k\}$ *is generated by Algorithm* 1. *Let* $x^*\in X^*$, *and choose* $k_0 := k_0(\bar{C}_2,\gamma\widehat{L}_4^2)\in\mathbb{N}, \phi\in(0,1)$ *such that*

$$\sum_{i\geq k_0}^{\infty}\frac{1}{N_i} \leq \frac{\phi}{\bar{C}_2\gamma^2\widehat{L}_4^2}, \quad (4.1)$$

then

$$\sup_{k\geq k_0} \big| \left\|x^k - x^*\right\| \big|_2^2 < \frac{\big| \left\|x^{k_0} - x^*\right\| \big|_2^2 + \frac{\phi C_2 \delta_4(x^*)^2}{\bar{C}_2 \widehat{L}_4^2}}{1-\phi}.$$

Proof. We have from $(A3)$ that there exists $k_0 \in \mathbb{N}_0$ satisfying (4.1). Let $d_i := \left\|x^i - x^*\right\|^2$ for $i \in \mathbb{N}_0$ and $k \geq k_0$ be given. Taking expectation in (3.11), making use of $\mathbb{E}[\mathbb{E}[\cdot|\widehat{\mathcal{F}}_k]|\mathcal{F}_k] = \mathbb{E}[\cdot|\mathcal{F}_k]$, and summing recursively from $i = k_0$ to $i = k-1$, we obtain

$$|d_k|_2^2 \leq |d_{k_0}|_2^2 + \bar{C}_2\gamma^2\widehat{L}_4^2 \sum_{i=k_0}^{k-1} \frac{|d_i|_2^2}{N_i} + C_2\gamma^2\delta_4^2(x^*) \sum_{i=k_0}^{k-1} \frac{1}{N_i}. \tag{4.2}$$

Let $t_\omega := \{k \geq k_0 \| d_k|_2 > \omega\}$. For any $\omega > 0$ and $t_\omega < \infty$, it follows from (4.1) and (4.2) that

$$\omega^2 < |d_{t_\omega}|_2^2 \leq |d_{k_0}|_2^2 + \bar{C}_2\gamma^2\widehat{L}_4^2 \sum_{i=k_0}^{t_\omega-1} \frac{|d_i|_2^2}{N_i} + C_2\gamma^2\delta_4^2(x^*) \sum_{i=k_0}^{t_\omega-1} \frac{1}{N_i} < |d_{k_0}|_2^2 + \phi\omega^2 + \frac{\phi C_2\delta_4^2(x^*)}{\bar{C}_2\widehat{L}_4^2}.$$

This implies that any threshold ω^2, which $\{|d_k|_2^2\}_{k\geq k_0}$ eventually exceeds, is bounded above by $\frac{|d_{k_0}|_2^2 + \frac{\phi C_2\delta_4^2(x^*)}{\bar{C}_2\widehat{L}_4^2}}{1-\phi}$. Therefore, $\{|d_k|_2^2\}_{k\geq k_0}$ is bounded and satisfies the statement of the lemma. □

Based on the above lemma, we next discuss the sublinear convergence rate in terms of the mean natural residual function and the oracle complexity.

Theorem 4.2 (sublinear convergence rate)**.** *Assume that* $(A0) - (A3)$ *hold. Let* $\rho := \frac{A_k(min\{\mu\theta,\gamma\})^2}{2|L(\xi)|_2^2}$ *and* $N_k := \mathcal{N}\lceil (k+\lambda)(\ln(k+\lambda))^{1+b}\rceil$ *for* $\mathcal{N} \in \mathbb{N}, b > 0, \lambda > 0$. *For any* $x^* \in X^*$, *if* $\sup_{k\geq 0} \big| \left\|x^k - x^*\right\| \big|_2^2 \leq M$ *for some* $M > 0$, *then for all* $k \in \mathbb{N}_0$,

$$\min_{i=0,\cdots,k} \mathbb{E}[H(x^i)^2] \leq \frac{1}{(k+1)\rho}\left(\left\|x^0 - x^*\right\|^2 + \frac{C_2\gamma^2\delta_4(x^*)^2 + \bar{C}_2\gamma^2\widehat{L}_4^2 M}{\mathcal{N} b\lceil \ln(\lambda-1)\rceil^b}\right).$$

Proof. From Remark 3.4, it is obvious that $\{N_k\}$ satisfies $(A3)$, and so Theorem 3.11 and Lemma 4.1 hold. It follows that the sequence $\{x^k\}$ is bounded in $\mathcal{L}^2$. Given $x^* \in X^*$ and $\sup_{k\geq 0} | \left\|x^k - x^*\right\| |_2^2 \leq M$ for some $M > 0$. Then, $\sup_k \mathbb{E}[\left\|x^k - x^*\right\| |^2] \leq M$. Taking expectation in (3.11), making use of $\mathbb{E}[\mathbb{E}[\cdot|\widehat{\mathcal{F}}_k]|\mathcal{F}_k] = \mathbb{E}[\cdot|\mathcal{F}_k]$ and summing recursively from $i = 0$ to $i = k$, we have

$$\rho\sum_{i=0}^{k} \mathbb{E}[H(x^i)^2] \leq \left\|x^0 - x^*\right\|^2 + (C_2\gamma^2\delta_4(x^*)^2 + \bar{C}_2\gamma^2\widehat{L}_4^2 M)\sum_{i=0}^{k}\frac{1}{N_i}.$$

Note that

$$\sum_{i=0}^{k}\frac{1}{N_i} \leq \sum_{i=0}^{\infty}\frac{1}{N_i} \leq \int_{-1}^{\infty}\frac{dq}{\mathcal{N}(q+\lambda)(\ln(q+\lambda))^{1+b}} = \frac{1}{\mathcal{N} b(\ln(\lambda-1))^b}$$

and $\min_{i=0,\cdots,k}\mathbb{E}[H(x^i)^2] \leq \frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}[H(x^i)^2]$, we obtain the conclusion immediately. □

Theorem 4.3 (oracle complexity)**.** *Let the assumptions of Theorem* 4.1 *hold and* $N_k := \mathcal{N}\lceil (k+\lambda)(\ln(k+\lambda))^{1+b}\rceil$ *for* $\mathcal{N} := \mathcal{O}(n), b > 0, \lambda > 0$. *For* $\tau > 0$, *Algorithm* 1 *achieves the tolerance*

$$\min_{i=0,\cdots,K} \mathbb{E}[H(x^i)^2] \leq \tau,$$

after $K := b^{-1}\mathcal{O}(\tau^{-1})$ *iterations and, with probability* 1, *an oracle complexity* $\sum_{i=0}^{K}(1+l_i)N_i$ *bounded above by*

$$b^{-2} \cdot \ln_{\frac{1}{\theta}} \left(\frac{\gamma \max_{i=0,\cdots,K} \widehat{L}_i}{\min\{\mu\theta, \gamma\}} \right) \cdot \lceil \ln(b^{-1}\tau^{-1}) \rceil^{1+b} \cdot \mathcal{O}(n\tau^{-2}),$$

where l_k *is the number of oracle calls used in the line search (3.1) at iteration* k *and* $\widehat{L}_k$ *is defined as in Lemma* 4.1. *Moreover, the mean oracle complexity* $\sum_{i=0}^{K}(1+\mathbb{E}[l_i])N_i$ *satisfies the same upper bound with* $\max_{i=0,\cdots,K} \widehat{L}_i$ *replaced by* L.

Proof. By Theorem 4.2, there exists a constant $\mathcal{M} > 0$ such that

$$\min_{i=0,\cdots,K} \mathbb{E}[H(x^i)^2] \leq \mathcal{M}n(\mathcal{N}bk)^{-1}, \ \ \forall k \in \mathbb{N}.$$

Thus, we have from $\tau > 0$ that $\min_{i=0,\cdots,K} \mathbb{E}[H(x^i)^2] \leq \tau$ after $K = \mathcal{O}(n\mathcal{N}^{-1}\tau^{-1})$ iterations. The total number of the oracle call is given by

$$\begin{aligned}
\sum_{i=0}^{K}(1+l_i)N_i &\lesssim \Big(\max_{i=0,\cdots,K} l_i\Big) \sum_{i=0}^{K} N_i (\ln i)^{1+b} \\
&\lesssim \Big(\max_{i=0,\cdots,K} l_i\Big) K^2 \mathcal{N}(\ln K)^{1+b} \\
&\lesssim \Big(\max_{i=0,\cdots,K} l_i\Big) \mathcal{N}^{-1} n^2 b^{-2} \tau^{-2} (\ln(n\mathcal{N}^{-1}b^{-1}\tau^{-1}))^{1+b} \qquad (4.3)
\end{aligned}$$

and $\min_{i=0,\cdots,K} \mathbb{E}[H(x^i)^2] \leq \tau$. From Lemma 3.5, we have $l_k \leq \log_{\frac{1}{\theta}}(\frac{\gamma \widehat{L}_k}{\min\{\mu\theta,\gamma\}})$. Hence, we can get the claimed bound on $\sum_{i=0}^{K}(1+l_i)N_i$ from (4.3) and $\mathcal{N} = \mathcal{O}(n)$ immediately.

On the other hand, the concavity of the mapping $z \longmapsto \log_{\frac{1}{\theta}}(z)$ and the Jensen's inequality imply that

$$\mathbb{E}[l_i] \leq \mathbb{E}[\log_{\frac{1}{\theta}} \Big(\frac{\gamma \widehat{L}_k}{\min\{\mu\theta, \gamma\}} \Big)] \leq \log_{\frac{1}{\theta}} \Big(\frac{\gamma L}{\min\{\mu\theta, \gamma\}} \Big),$$

where the last inequality follows from $\mathbb{E}[\widehat{L}_k] = L$ by the definitions of $\{\widehat{L}_k, L\}$ and $(A3)$. This together with (4.3) and $\mathcal{N} = \mathcal{O}(n)$ imply the claimed bound on the mean oracle complexity $\sum_{i=0}^{K}(1+\mathbb{E}[l_i])N_i$. ☐

Remark 4.4. By Lemma 4.1, the constant M in Theorem 4.2 can be estimated by

$$M \leq \frac{\max_{k=0,\cdots,k_0} \big| \left\| x^k - x^* \right\| \big|_2^2 + \frac{\phi C_2 \delta_4 (x^*)^2}{\bar{C}_2 \widehat{L}_4^2}}{1-\phi}.$$

Moreover, the number k_0 in the above inequality can be estimated from (4.1) by

$$k_0 \geq exp \left(\sqrt[b]{\frac{\bar{C}_2 \gamma^2 \widehat{L}_4^2}{\phi b \mathcal{N}}} \right) - \lambda + 1.$$

To obtain the line convergence rate of the Algorithm 1, we need the following assumption.

Assumption 4.1 ($A4$)**.** The SMVI satisfy the bounded metric subregularity condition at any $\bar{x} \in X^*$, namely, for any compact set V satisfying $X^* \subseteq V$, there exsits a constant $\kappa > 0$ such that

$$dist(x, X^*) \leq \kappa dist(0, F(x^*) + \partial g(x)), \quad \forall x \in V.$$

According to Proposition 3 in [39], the bounded metric subregularity condition is equivalent to the bounded proximal error bound condition at $\bar{x} \in X^*$, that is, for any $\lambda > 0$ and compact set V satisfying $X^* \subseteq V$, there exists a constant $\kappa > 0$ such that

$$dist(x, X^*) \leq \kappa \left\| x - \text{prox}_{\lambda g}[x - \lambda F(x^*)] \right\|, \quad \forall x \in V.$$

Next, we give an example which satisfies the bounded proximal error bound condition.

Example 4.1 (Stochastic Networked Nash-Cournot game)**.** Consider a networked Nash-Cournot game with uncertain data in [19, 35, 41, 33], in which the cost-minimizing agents compete in quantity levels when facing with a price function associated with aggregate output. Assume that there are $\mathcal{I}$ firms that compete over a network of $\mathcal{J}$ nodes in supplying a homogeneous product in a non-cooperative fashion. The level of sales of firm $i \in [\mathcal{I}]$ at node $j \in [\mathcal{J}]$ is denoted by x_{ij}, and the firm is characterized by a random linear production cost function $c_i(x_i, \xi_i) = (a_i + \xi_i) \sum_{j \in [\mathcal{J}]} x_{ij}$ for some parameter $a_i > 0$, where ξ_i is a mean-zero random variable. Assume that the price at node j, denoted by $P_j(\sum_{i \in [\mathcal{I}]} x_{ij}, \eta_j)$ is a stochastic linear function corrupted by noise $P_j(\sum_{i \in [\mathcal{I}]} x_{ij}, \eta_j) = d_j + \eta_j - b_j \sum_{i \in [\mathcal{I}]} q_{ij}$, where d_j indicates the price when the production is zero, b_j represents the slope of the inverse demand function, and η_j is a zero-mean random disturbance. For simplicity, we assume the transportation costs to be zero. Except the nonnegativity constraints on x_{ij}, we suppose that firm $i's$ production at node j is capacitated by cap_{ij}. Therefore, firm $i's$ optimization problem is given by

$$\min \mathbb{E}\Big[f_i(x, \xi, \eta)] = \mathbb{E}[c_i(x_i, \xi_i) - \sum_{j \in [\mathcal{J}]} P_j(\sum_{i \in [\mathcal{I}]} x_{ij}, \eta_j) x_{ij}\Big]$$

$$s.t.\ x_i \in X_i = \{x_i \in \mathbb{R}^{\mathcal{J}} | x_i \geq 0, x_{ij} \leq \text{cap}_{ij}\}.$$

Under some dominated conditions, we may interchange the orders of expectation and derivative so that the above stochastic Nash-Cournot game may be transformed into the SMVI, in which $g(x) = \sigma_X(x)$ with $X = \Pi_{i=1}^{\mathcal{I}} X_i$ and $F(x^*) = (F_1(x^*), \cdots, F_I(x^*)), F_i(x^*) = \mathbb{E}[\partial_{x_i}, f_i(x, \xi, \eta)]$. It can be proved that the bounded proximal error bound at any $\bar{x} \in X^*$ with $X^* = \Pi_{i \in \mathcal{I}} X_i^*$ is satisfied. In fact, by letting $g_i(x_i) = \sigma_{X_i}(x_i)$ with $\sigma_{X_i}(x_i) = 0$ if $x_i \in X_i$ and ∞ otherwise, the stochastic Nash-Cournot game is equivalent to the following generalized equations:

$$0 \in F_i(x^*) + \partial g_i(x_i), \ \ i \in \mathcal{I}.$$

Since X_i is a polyhedral set and $\text{diag}(b)$ is a semidefinite diagonal matrix, $\partial g_i(x_i)$ is a polyhedral multifunction. Hence, by [39], the bounded proximal error bound at $\bar{x}$ for any $\bar{x} \in X^*$ is satisfied, i.e., for any compact set V containing $\bar{x}$, there exists κ such that

$$dist(x, X^*) \leq \kappa \left\| x - \text{prox}_{\lambda g}[x - \lambda F(x^*)] \right\|, \quad \forall x \in V.$$

Theorem 4.5 (linear convergence rate)**.** *Let the sequence $\{x^k\}$ be generated by Algorithm* 1 *and $K > 0$ be a given integer. Assume that there exists a compact set V such that $X^* \subseteq V$ and $x^k \in V$ $(0 \leq k \leq K)$. Consider (A0)-(A4) and the following conditions:*

(i) *There exists* $\bar{D} > 0$ *such that* $\mathbb{E}[dist^2(x^0, X^*)] < \bar{D}$;

(ii) $0 < \varphi := 1 - \frac{\rho}{\kappa^2} + \frac{\bar{C}_2\gamma^2\widehat{L}_4^2}{N_0} \le 1$;

(iii) *For any* $k \in \mathbb{N}_0$, *we have* $N_k \ge N_0$ *and* $N_k \ge \lceil \frac{\varrho_K}{\varphi^k} \rceil$, *where* $\varrho_K := \frac{\sum_{k=1}^K N_k - K}{\sum_{k=1}^K \frac{1}{\varphi^k}}$.

Then there exsits $\mathcal{D} > 0$ *such that*

$$\mathbb{E}[dist^2(x^{k+1}, X^*)] \le \varphi^k \left(\varphi\bar{D} + (K+1)\frac{\mathcal{D}}{\varrho_K} \right), \quad \forall k \in K.$$

Proof. Let $\bar{x}^k := \text{prox}_g[x^k]$. By $x^k \in \mathcal{F}_k$ and the continuity of the proximal mapping, $\bar{x}^k \in \mathcal{F}_k$. From (3.11), $dist(x^k, X^*) = \inf \|x^k - \bar{x}^k\|$ and $N_0 \le N_k (\forall k \in \mathbb{N}_0)$, we have

$$\begin{aligned}
\mathbb{E}[dist^2(x^{k+1}, X^*)|\mathcal{F}_k] &\le \mathbb{E}[\|x^{k+1} - \bar{x}^k\|^2 |\mathcal{F}_k] \\
&\le \left(1 + \frac{\bar{C}_2\gamma^2\widehat{L}_4^2}{N_0}\right) \|x^k - \bar{x}^k\|^2 - \rho H(x^k)^2 + \frac{C_2\gamma^2\delta_4^2(\bar{x}^k)}{N_k} \\
&\le \left(1 - \frac{\rho}{\kappa^2} + \frac{\bar{C}_2\gamma^2\widehat{L}_4^2}{N_0}\right) dist^2(x^k, X^*) + \frac{C_2\gamma^2\delta_4^2(\bar{x}^k)}{N_k}. \qquad (4.4)
\end{aligned}$$

Taking $\mathbb{E}[\cdot|\mathcal{F}_k]$ in (4.4), we obtain

$$\mathbb{E}[dist^2(x^{k+1}, X^*)] \le \varphi\mathbb{E}[dist^2(x^k, X^*)] + \frac{C_2\gamma^2\delta_4^2(\bar{x}^k)}{N_k}.$$

Note that $\delta_4(\cdot)$ is continuous, $\bar{x}^k \in X^*$ and X^* is bounded, we obtain that there exsits $\mathcal{D} > 0$ such that $C_2\gamma^2\delta_4^2(\bar{x}^k) \le \mathcal{D}$. From $\varphi \in (0,1)$, we have

$$\begin{aligned}
\mathbb{E}[dist^2(x^{k+1}, X^*)] &\le \varphi\mathbb{E}[dist^2(x^k, X^*)] + \frac{\mathcal{D}}{N_k} \\
&\le \varphi^{k+1}\mathbb{E}[dist^2(x^0, X^*)] + \varphi^k\frac{\mathcal{D}}{N_0} + \varphi^{k-1}\frac{\mathcal{D}}{N_1} + \cdots + \frac{\mathcal{D}}{N_k} \\
&\le \varphi^{k+1}\mathbb{E}[dist^2(x^0, X^*)] + \varphi^k\frac{\mathcal{D}}{\varrho_K} + \varphi^k\frac{\mathcal{D}}{\varrho_K} + \cdots + \varphi^k\frac{\mathcal{D}}{\varrho_K} \\
&\le \varphi^k \left(\varphi\bar{D} + (K+1)\frac{\mathcal{D}}{\varrho_K} \right).
\end{aligned}$$

This completes the proof. □

5 Numerical Results

5.1 Numerical experiments

In this section, we give some numerical examples to illustrate the efficiency of Algorithm 1 by comparing with several related algorithms such as Algorithms VBPE and VBPFB in [32], Algorithm VBMBF in [38] and Algorithm EGLS in [13] according to the CPU time and the empirical errors. All algorithms were coded in MATLAB R2019a and run the same computer with Windows 10 system, AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx 2.10 GHz and 16.0 GB. We use *randn* and *unifrnd* in Matlab 2019a to generate i.i.d samples from Ξ and terminate algorithms when the total number of iterations reaches 8000. The numerical results are listed on the following, where 'iter' and 'time' denote the number of iterations and the CPU time in seconds, respectively. In addition, we define the empirical error:$=\frac{x^{k+1}-x^*}{x^*}$ and the tolerance $\varepsilon := 10^{-4}$. Note that dom$g$ is a nonempty, closed and convex set in this section. See Figs. $1-4$ and Table 5.1 for more details about the numerical results.

The parameters are chosen as follows:
Algorithm 1: $\gamma = 0.99, \theta = 0.4, \beta = 0.7, \mu = 0.2, N_k = (k+100)(\ln(k+100))^2$.
VBPE: $\alpha_k = 0.1, N_k = \lceil 0.7^{-(k+1)} \rceil$.
VBPFB: $\alpha_k = 0.1, N_k = \lceil 0.7^{-(k+1)} \rceil$.
VBMBF: $\gamma = 0.99, \theta = 0.4, \mu = 0.2, N_k = (k+100)(\ln(k+100))^2$.
EGLS: $\gamma = 0.99, \theta = 0.4, \beta = 0.1, \lambda = 0.2, N_k = \lceil 0.7^{-(k+1)} \rceil$.

Example 5.1. Consider the following SMVI in which ξ is uniformly distributed on $\Xi = [0,1]$, $x \in X = [0,4]\times[0,4]\times[0,4]$ and $f : \mathbb{R}^3 \times \Xi \to \mathbb{R}^3$, $g : \mathbb{R}^3 \to \mathbb{R}$ are given by

$$f(x,\xi) := \begin{pmatrix} x_1 - \xi x_2 + 3 - 2\xi \\ -\xi x_1 + 2x_2 + \xi x_3 - 2 - \xi \\ \xi x_2 + 3x_3 - 3 - \xi \end{pmatrix},$$

and

$$g(x) := \frac{1}{2}x_1^2 + x_2^2 + x_3^2 - x_1x_2 + x_2x_3 + x_1 - 3x_2 - 3x_3 + 6.$$

Numerical results are shown in Fig.1.

Example 5.2. In Example 1, $g(x)$ has been replaced by $g(x) := x_1^2 - x_1x_2 - 2x_2x_3 + x_1 + 2x_2 + 2x_3$. Numerical results are shown in Fig.2.

Table 5.1: Comparison of Algorithm 1 and VBMBF.

	$\epsilon = 10^{-4}$		$\epsilon = 10^{-8}$		$\epsilon = 10^{-15}$	
	Algorithm 1	VBMBF	Algorithm 1	VBMBF	Algorithm 1	VBMBF
Ex.1(X=50)						
time	0.1983	0.5382	0.4044	0.7960	0.4210	0.7734
iter	373	1000	1000	1000	1000	1000
Ex.1(X=100)						
time	0.6216	1.5663	0.8955	1.3509	0.9431	1.3397
iter	729	1000	1000	1000	1000	1000
Ex.1(X=200)						
time	1.8561	3.5202	2.1475	3.2118	2.2866	3.3194
iter	915	1000	1000	1000	1000	1000
Ex.2(X=50)						
time	0.2740	0.4856	0.3983	0.6780	0.4531	0.7578
iter	1000	1000	1000	1000	1000	1000
Ex.2(X=100)						
time	0.7995	1.4846	0.8373	1.4661	0.9050	1.3753
iter	1000	1000	1000	1000	1000	1000
Ex.2(X=200)						
time	1.9069	2.9232	2.0519	3.0930	2.2227	2.9678
iter	1000	1000	1000	1000	1000	1000

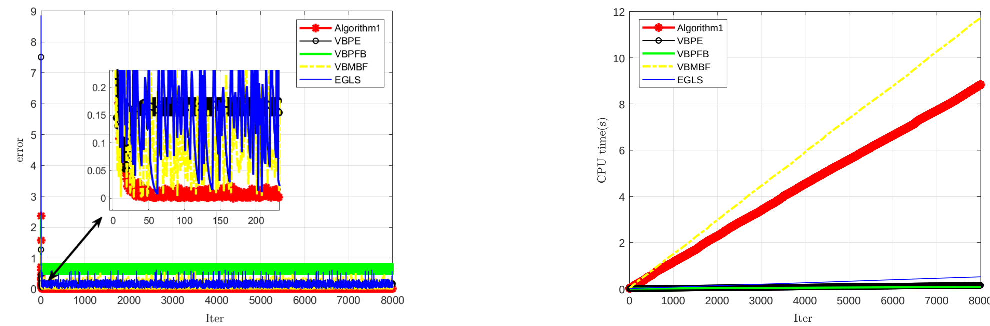

Fig.1 Results for Example 5.1.

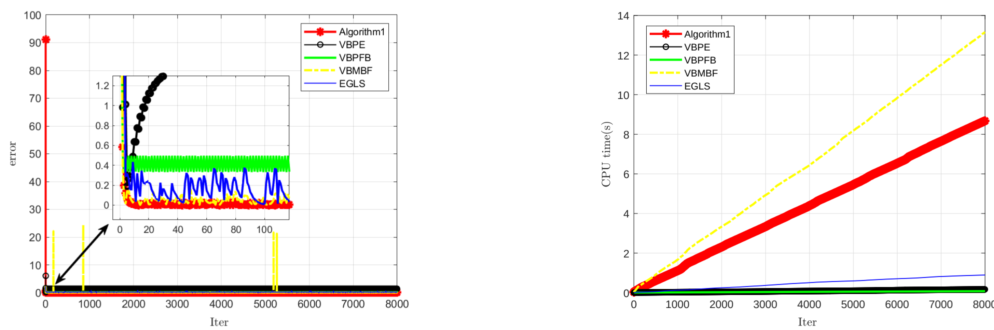

Fig.2 Results for Example 5.2.

In Figs.1-2 and Table 5.1, we show the performances of all algorithms in solving the quadratic function problems presented in Examples 5.1 and 5.2. It is observed that the empirical error solved by Algorithms VBPE, VBPFB, VBMBF and EGLS vary with the same decreasing trend as Algorithm 1, while the Algorithm VBMBF spends the most CPU time among them.

Note that in the special case where $P_C(x) = \text{prox}_{\lambda\sigma_C}(x)$, Algorithm 1 degenerates to Algorithm VBMBF proposed in [38] for SVI. By comparing Algorithm 1 with Algorithm VBMBF in Examples 5.1 and 5.2, we have from Figs.1-2 and Table 5.1 that Algorithm 1 is more competitive both in the CPU time (Algorithm 1 saves nearly 30% CPU time) and the empirical error. In particular, Algorithm 1 possesses better performances on stability while Algorithm VBMBF fluctuates greatly in terms of the empirical error. Moreover, Algorithm VBMBF outperforms Algorithms VBPE, VBPFB and EGLS with respect to the empirical error, which implies that our algorithm performs well even in the simplest case, namely, $P_C(x) = \text{prox}_{\lambda\sigma_C}(x)$.

In a word, the results shown in Figs.1-2 and Table 5.1 indicate that Algorithm 1 is better than other algorithms (although Algorithms VBPE, VBPFB and EGLS take less CPU time, their empirical errors decrease slowly and far away from 0, which means that the solutions solved by them may be rougher than the ones solved by Algorithm 1).

Example 5.3. Consider the following SMVI: Find $x^* \in X$ such that

$$\mathbb{E}[A(\xi)x^* + b(\xi)]^T(x - x^*) + g(x) - g(x^*) \geq 0, \ \forall x \in X,$$

where $g(x) := \lambda \|x\|_1$, λ is the regularization parameter, $A(\xi) \in \mathbb{R}^{n\times n}$ is a random matrix, $b(\xi)$ is a random vector, both $A(\xi)$ and $b(\xi)$ are unknown ahead of time. Given $x^* = (0, 0, \cdots, 0)^T$. In our test, we generate the matrix $A(\xi)$ by $A(\xi) = \frac{M+M^T}{2}$, where all elements of $M \in \mathbb{R}^{n\times n}$ and $b(\xi) \in \mathbb{R}^n$ are both uniformly distributed random samples generated from [0,1], and the dimension of the tested problem is 500. Numerical results are shown in Fig.3.

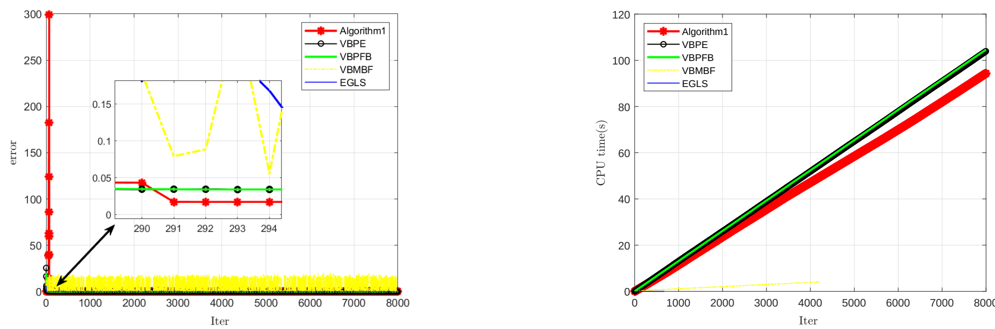

Fig.3 Results for Example 5.3.

Example 5.4. This calculation example is derived from $g(x) := \lambda \|x\|_2$ instead of $g(x)$ in Example 5.3. The selections of other functions and parameters are consistent with Example 5.3. Numerical results are shown in Fig.4.

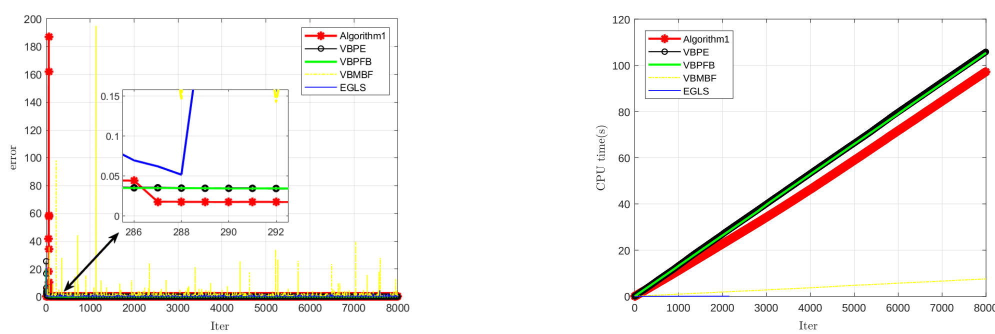

Fig.4 Results for Example 5.4.

The numerical results are shown in Figs.3-4. In Examples 5.3 and 5.4, we consider the general SMVI problem. As is shown in the figures, Algorithms VBMBF and EGLS take less running time than other variance-based algorithms, but the empirical error solved by Algorithm 1 is hardly decayed among all algorithms. Besides, Algorithm 1 outperforms Algorithms VBPE and VBPFB even if they take approximately the same CPU time. All in all, for the tested problems, Algorithm 1 is better than Algorithm VBPE, VBPFB, VBMBF and EGLS in the calculation accuracy, while Algorithms VBMBF and EGLS are slightly competitive with Algorithm 1 in calculation speed.

5.2 Application of Algorithm 1 to the stochastic Nash game

In this subsection, we apply the Algorithm 1 to solve the stochastic Nash game discussed in Example 4.1. In our test, we consider a Cournot game with $\mathcal{I} = 5, 10, 20$ firms and $\mathcal{J} = 10$

markets. For any $i \in \mathcal{I}$ and $j \in \mathcal{J}$, we set $cap_{ij} = 2, d_j \sim U(40, 50), b_j \sim U(1, 2), a_i \sim U(3, 5)$, where $U(\underline{u}, \overline{u})$ denotes the uniform distribution on $[\underline{u}, \overline{u}]$ and $\underline{u} < \overline{u}$. Moreover, let $\xi_i \sim U(-a_i/5, a_i/5), \eta_j \sim U(-b_i/5, b_i/5)$ and the initial point $x^0 = (1, ..., 1)^T$. The specific numerical results are shown in table 5.2, where the relative error:$=\frac{\|x^K - x^*\|}{\|x^*\|}$.

Table 5.2: The numerical results of stochastic Nash game

The number of the firm	Max Iterations	Relative error	CPU time (s)
I=5	100	3.8210e-1	1.1754e+1
	500	1.0700e-2	5.6553e+1
	1000	3.7000e-3	1.1183e+2
	2000	2.9000e-3	2.2700e+2
I=10	100	8.3800e-2	2.5801e+0
	500	6.9200e-2	7.8577e+1
	1000	1.8900e-2	1.6774e+2
	2000	2.6000e-3	3.6233e+2
I=20	100	1.6950e-1	4.2369e+1
	500	7.3800e-2	2.0949e+2
	1000	2.9300e-2	4.9200e+2
	2000	8.1000e-3	8.7351e+2

From Table 5.2, we can see that the number of the firm and the maximum iteration may slow down the values of relative error. On the contrary, the CPU time may increase with the increase of the number of the firm and the maximum iteration.

6 Conclusion

In this paper, we introduce a modified proximal backward-forward algorithm with variance reduction for stochastic mixed variational inequalities. The new algorithm uses the line search scheme which is not necessarily to know the Lipschitz constant and requires only one evaluation of the proximal mapping per iteration. Besides, we obtain the asymptotic convergence, the sublinear convergence rate and the optimal oracle complexity in terms of the mean natural residual function. We also discuss the linear convergence rate with finite computational budget under the assumption of the bounded proximal error bound. Finally, some numerical results are obtained to show the superiority of the new algorithm.

References

[1] A. Alacaoglu, Y. Malitsky and V. Cevher, Forward-reflected-backward method with variance reduction, *Comput. Optim. Appl.* 80 (2021) 321–346.

[2] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer-Verlag, New York, 2011.

[3] R.I. Bot, P. Mertikopoulos, M. Staudigl and P.T. Vuong, Minibatch forward-backward-forward methods for solving stochastic variational inequalities, *Stoch. Syst.* 11 (2021) 112–139.

[4] D.L. Burkholder, B.J. Davis and R.F. Gundy, Integral inequalities for convex functions of operators on martingales, in: *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1972, pp. 223–240.

[5] L. Chen, G.H. Lin and X.M. Yang, Expected residual minimization method for stochastic mixed variational inequality problems, *Pacific J. Optim.* 14 (2018) 703–722.

[6] L. Chen, G.H. Lin and X.M. Yang, Sample average approximation methods based on regularized gap functions for stochastic mixed variational inequality problems, *J. Nonl. Convex Anal.* 20 (2019) 149–164.

[7] S. Cui and U.V. Shanbhag, On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems, in: *IEEE 55th Conference on Decision and Control*, 2016, pp. 4510–4515.

[8] S.D. Flam, Games and cost of change, *Ann. Operat. Res.* 301 (2021) 107–119.

[9] F. Facchinei and J.S. Pang, *Finite-dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.

[10] D. Goeleven, *Complementarity and Variational Inequalities in Electronics*, Academic Press, London, 2017.

[11] S.M. Grad and F. Lara, Solving mixed variational inequalities beyond convexity, *J. Optim. Theory Appl.* 190 (2021) 565–580.

[12] W.M. Han and B. Reddy, On the finite element method for mixed variational inequalities arising in elastoplasticity, *SIAM J. Numer. Anal.* 32 (1995) 1778–1807.

[13] A.N. Iusem, A. Jofré, R.I. Oliveira and P. Thompson, Extragradient method with variance reduction for stochastic variational inequalities, *SIAM J. Optim.* 27 (2017) 1143–1153.

[14] A.N. Iusem, A. Jofré, R.I. Oliveira and P. Thompson, Variance-based extragradient methods with line search for stochastic variational inequalities, *SIAM J. Optim.* 29 (2019) 175–206.

[15] A.N. Iusem, A. Jofré and P. Thompson, Incremental constraint projection methods for monotone stochastic variational inequalities, *Math. Oper. Res.* 44 (2019) 236–263.

[16] H.Y. Jiang and H.F. Xu, Stochastic approximation approaches to the stochastic variational inequality problem, *IEEE Trans. Autom. Control.* 53 (2008) 1462–1475.

[17] A. Kannan and U.V. Shanbhag, Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants, *Comput. Optim. Appl.* 74 (2019) 779–820.

[18] A. Kannan and U.V. Shanbhag, The pseudomonotone stochastic variational inequality problem: Analytical statements and stochastic extragradient schemes, in: *In 2014 American Control Conference*, 2014, pp. 2930–2935.

[19] J.F. Koshal, A. Nedić and U.V. Shanbhag, Regularized iterative stochastic approximation methods for stochastic variational inequality problems, *IEEE Trans. Autom. Control.* 58 (2013) 594–609.

[20] I. Konnov and E.O. Volotskaya, Mixed variational inequalities and economic equilibrium problems, *J. Appl. Math.* 6 (2002) 289–314.

[21] J.L. Lei and U.V. Shanbhag, Distributed variable sample-size gradient-response and best-response schemes for stochastic Nash equilibrium problems over graphs, *SIAM J. Optim.* 32 (2022) 573–603.

[22] Y. Malitsky, Golden ratio algorithms for variational inequalities, *Math. Program.* 184 (2020) 383–410.

[23] Y. Malitsky, Proximal extrapolated gradient methods for variational inequalities, *Optim. Math. Soft.* 33 (2018) 140–164.

[24] K. Mishchenko, D. Kovalev, E. Shulgin, Y. Malitsky and P. Richtárik, Revisiting stochastic extragradient, in: *In the 23rd International Conference on Artificial Intelligence and Statistics*, vol. 108, 2020, pp. 4573–4582.

[25] B. Polyak, *Introduction to Optimization*, Optim. Soft. Press, New York, 1987.

[26] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Stat.* 22 (1951) 400–407.

[27] R.T. Rockafellar and R.J. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1988.

[28] H.L. Sun and X.J. Chen, Two-stage stochastic variational inequalities: theory, algorithms and applications, *J. Oper. Res. Soc. China.* 9 (2021) 1–32.

[29] G.J. Tang, Z.P. Wan and N.J. Huang, Strong convergence of a projection-type method for mixed variational inequalities in Hilbert spaces, *Numer. Func. Anal. Optim.* 39 (2018) 1103-1119.

[30] G.J. Tang, M. Zhu and H.W. Liu, A new extragradient-type method for mixed variational inequalities, *Oper. Res. Lett.* 43 (2015) 567–572.

[31] M. Wang and D.P. Bertsekas, Incremental constraint projection methods for variational inequalities, *Math. Program.* 150 (2015) 321–363.

[32] Z.P. Yang and G.H. Lin, Two fast variance-reduced proximal gradient algorithms for SMVIPs-stochastic mixed variational inequality problems with suitable applications to stochastic network games and traffic assignment problems, *J. Comput. Appl. Math.* 408 (2022): 114132.

[33] Z.P. Yang and G.H. Lin, Variance-based single-call proximal extragradient algorithms for stochastic mixed variational inequalities, *J. Optim. Theory Appl.* 190 (2021) 393–427.

[34] F. Yousefian, A. Nedić and U.V. Shanbhag, Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and Nash games, *IEEE Trans. Autom. Control.* 61 (2016) 1753–1766.

[35] F. Yousefian, A. Nedić and U.V. Shanbhag, On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems, *Math. Program.* 165 (2017) 391–431.

[36] F. Yousefian, A. Nedić and U.V. Shanbhag, Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems, in: *In 53rd IEEE Conference on Decision and Control*, 2014, pp. 5831–5836.

[37] F. Yousefian, A. Nedić and U.V. Shanbhag, On stochastic mirror-prox algorithms for stochastic Cartesian variational inequalities: Randomized block coordinate and optimal averaging schemes, *Set-Valued Var. Anal.* 26 (2018) 789–819.

[38] Z.P. Yang, Y.L. Wang and G.H. Lin, Variance-based modified backward-forward algorithm with line search for stochastic variational inequality problems and its applications, *Asia-Pac. J. Oper. Res.* 37 (2020): 2050011.

[39] J.J. Ye, X.M. Yuan, S.Z. Zeng and J. Zhang, Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems, *Set-Valued Var. Anal.* 29 (2021) 803–837.

[40] X.M. Yang, Y. Zhao, G.H. Lin, Deterministic bicriteria model for stochastic variational inequalities, *J. Oper. Res. Soc. China.* 6 (2018) 507–527.

[41] Z.P. Yang, J. Zhang, Y.L. Wang and G.H. Lin, Variance-based subgradient extragradient method for stochastic variational inequality problems, *J. Sci. Comput.* 89 (2021) 1–28.

[42] X.J. Zhang, X.W. Du, Z.P. Yang and G.H. Lin, An infeasible stochastic approximation and projection algorithm for stochastic variational inequalities, *J. Optim. Theory Appl.* 183 (2019) 1053–1076.

[43] R.Y. Zhong and N.J. Huang, Stability analysis for minty mixed variational inequality in reflexive Banach spaces, *J. Optim. Theory Appl.* 147 (2010) 454–472.

Manuscript received 30 September 2021
revised 30 January 2022, 4 March 2022
accepted for publication 30 April 2022

Yue-Hong He
College of Mathematics and Statistics
Chongqing Technology and Business University
Chongqing 400067, China
E-mail address: heyuehong1111@163.com

Xian-Jun Long
College of Mathematics and Statistics
Chongqing Technology and Business University
Chongqing 400067, China
E-mail address: xianjunlong@ctbu.edu.cn