Math. Program., Ser. B (2017) 165:113–149
DOI 10.1007/s10107-017-1161-4

FULL LENGTH PAPER

Accelerated schemes for a class of variational inequalities

Yunmei Chen[1] · **Guanghui Lan**[2] · **Yuyuan Ouyang**[3]

Received: 19 August 2014 / Accepted: 10 May 2017 / Published online: 1 June 2017

Abstract We propose a novel stochastic method, namely the stochastic accelerated mirror-prox (SAMP) method, for solving a class of monotone stochastic variational inequalities (SVI). The main idea of the proposed algorithm is to incorporate a multi-step acceleration scheme into the stochastic mirror-prox method. The developed SAMP method computes weak solutions with the optimal iteration complexity for SVIs. In particular, if the operator in SVI consists of the stochastic gradient of a smooth function, the iteration complexity of the SAMP method can be accelerated in terms of their dependence on the Lipschitz constant of the smooth function. For SVIs with bounded feasible sets, the bound of the iteration complexity of the SAMP method depends on the diameter of the feasible set. For unbounded SVIs, we adopt the modified gap function introduced by Monteiro and Svaiter for solving monotone inclusion, and show that the iteration complexity of the SAMP method depends on the distance

Yunmei Chen is partially supported by NSF Grants DMS-1115568, IIP-1237814 and DMS-1319050. Guanghui Lan is partially supported by NSF Grants CMMI-1637473, CMMI-1637474, DMS-1319050 and ONR Grant N00014-16-1-2802. Part of the research was done while Yuyuan Ouyang was a Ph.D. student at the Department of Mathematics, University of Florida, and Yuyuan Ouyang is partially supported by AFRL Mathematical Modeling Optimization Institute.

✉ Guanghui Lan
George.lan@isye.gatech.edu

Yunmei Chen
yun@math.ufl.edu

Yuyuan Ouyang
yuyuano@clemson.edu

1 Department of Mathematics, University of Florida, Gainesville, FL, USA

2 School of Industrial and System Engineering, Georgia Institute of Technology, Atlanta, Georgia

3 Department of Mathematical Sciences, Clemson University, Clemson, SC, USA

from the initial point to the set of strong solutions. It is worth noting that our study also significantly improves a few existing complexity results for solving deterministic variational inequality problems. We demonstrate the advantages of the SAMP method over some existing algorithms through our preliminary numerical experiments.

Keywords Stochastic variational inequalities · Stochastic programming · Mirror-prox method · Extragradient method

Mathematics Subject Classification 90C25 · 90C15 · 62L20 · 68Q25

1 Introduction

Let $\mathcal{E}$ be a finite dimensional vector space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$ (not necessarily induced by $\langle\cdot,\cdot\rangle$), and Z be a non-empty closed convex set in $\mathcal{E}$. Our problem of interest is to find an $u^* \in Z$ that solves the following monotone stochastic variational inequality (SVI) problem:

$$\langle \mathbb{E}_{\xi,\zeta}[\mathcal{F}(u;\xi,\zeta)], u^* - u\rangle \le 0, \quad \forall u \in Z. \tag{1}$$

Here, the expectation is taken with respect to the random vectors ξ and ζ whose distributions are supported on $\Xi \subseteq \mathbb{R}^d$ and $\Xi' \subseteq \mathbb{R}^{d'}$, respectively, and $\mathcal{F}$ is given by the summation of three components with different structural properties, i.e.,

$$\mathcal{F}(u;\xi,\zeta) = \mathcal{G}(u;\xi) + \mathcal{H}(u;\zeta) + J'(u), \quad \forall u \in Z. \tag{2}$$

In particular, we assume that $J'(u) \in \partial J(u)$ is a subgradient of a relatively simple and convex function J (see (10) below), $\mathcal{H}(u;\zeta)$ is an unbiased estimator of a monotone and Lipschitz continuous operator H such that $\mathbb{E}_\zeta[\mathcal{H}(u;\zeta)] = H(u)$,

$$\langle H(w) - H(v), w - v\rangle \ge 0, \text{ and } \|H(w) - H(v)\|_* \le M\|w - v\|, \quad \forall w, v \in Z, \tag{3}$$

where $\|\cdot\|_*$ denotes the conjugate norm of $\|\cdot\|$. Moreover, we assume that $\mathcal{G}(u;\xi)$ is an unbiased estimator of the gradient for a convex and continuously differentiable function G such that $\mathbb{E}_\xi[\mathcal{G}(u;\xi)] = \nabla G(u)$ and

$$0 \le G(w) - G(v) - \langle \nabla G(v), w - v\rangle \le \frac{L}{2}\|w - v\|^2, \quad \forall w,\ v \in Z. \tag{4}$$

Observe that u^* given by (1) is often called a weak solution for SVI. A related notion is a strong SVI solution. More specifically, letting

$$F(u) := \mathbb{E}_{\xi,\zeta}[\mathcal{F}(u;\xi,\zeta)] = \nabla G(u) + H(u) + J'(u), \tag{5}$$

we say that u^* is a strong SVI solution if it satisfies

$$\langle F(u^*), u^* - u\rangle \leq 0, \quad \forall u \in Z. \tag{6}$$

It should be noted that the operator F above might not be continuous. Problems (1) and (6) are also known as the Minty variational inequality and the Stampacchia variational inequality respectively, due to their origin [16,27]. For any monotone operator F, it is well-known that strong solutions defined in (6) are also weak solutions in (1), and the reverse is also true under mild assumptions (e.g., when F is continuous). For example, for F in (5), if $J = 0$, then the weak and strong solutions in (1) and (6) are equivalent. For the sake of notational convenience, we use $SVI(Z; G, H, J)$ or simply $SVI(Z; F)$ to denote problem (1).

SVIs have recently found many applications, especially in data analysis. To motivate our discussion, let us mention one widespread machine learning model which helps to represent massive data in a compact way [17]. Consider a set of observed data $S = \{(x_i, y_i)\}_{i=1}^m$, drawn at random from an unknown distribution $\mathcal{D}$ on $X \times Y$. We would like to find a function $y = \mathcal{Y}(x, \theta)$, parameterized by $\theta \in \Theta$, to describe the relation between x and y. To this end, we can solve different problems of the form (e.g., [4,45,49,52])

$$\min_{\theta\in\Theta} \mathbb{E}[\mathcal{L}(\mathcal{Y}(x, \theta), y)] + r(\theta), \tag{7}$$

where $\mathcal{L}$ denotes a loss function, r is a regularization to enforce certain structures of the generated solutions, and the expectation is taken w.r.t. the random vector (x, y). While one can directly solve (7) as a stochastic optimization problem, the SVI in (1) provides us a unified model to study different subclasses of problems given in the form of (7), which include but not limited to the following cases when: (a) $\mathcal{L}$ is a smooth convex function (see [24]); (b) either $\mathcal{L}$ or r are nonsmooth, but admitting a saddle point reformulation (see [10]); (c) the feasible set contains linear or nonlinear functional constraints; and (d) θ has to satisfy the optimality condition for another optimization problem. Moreover, the SVI in (1) can potentially be used to solve a wider class of stochastic equilibrium and complementarity problems whose operators are given in the form of expectation (see for instance the survey [26] and the references therein).

In spite of the modeling power of SVIs and many different algorithms that have been developed for solving deterministic VIs [9,20,22,28,32,35,37,40,43,44], to compute the solutions of SVIs still seems to be challenging. A basic difficulty to solve (1) is that the expectation function in (1) cannot be computed efficiently within high accuracy, especially when the dimension of the random vector (ξ, ζ) is large. Hence, the algorithms for solving deterministic VIs are not directly applicable to SVIs in general. This paper focuses on Monte-carlo sampling (or scenario generation) based approaches for solving SVIs. In particular, we assume that there exist *stochastic oracles* $\mathcal{SO}_G$ and $\mathcal{SO}_H$ that provide random samples of $\mathcal{G}(u; \xi)$ and $\mathcal{H}(u; \xi)$ for any test point $u \in Z$. At the i-th call of $\mathcal{SO}_G$ and $\mathcal{SO}_H$ with input $z \in Z$, the oracles $\mathcal{SO}_G$ and $\mathcal{SO}_H$ output stochastic information $\mathcal{G}(z; \xi_i)$ and $\mathcal{H}(z; \zeta_i)$ respectively, such that

A1

$$\mathbb{E}\left[\|\mathcal{G}(u;\xi_i)-\nabla G(u)\|_*^2\right]\le\sigma_G^2,\ \mathbb{E}\left[\|\mathcal{H}(u;\zeta_i)-H(u)\|_*^2\right]\le\sigma_H^2,$$

for some $\sigma_G, \sigma_H \ge 0$, where ξ_i and ζ_i are independently distributed random samples. For the sake of notational convenience, throughout this paper we also denote

$$\sigma := \sqrt{\sigma_G^2+\sigma_H^2}. \tag{8}$$

Assumption A1 basically implies that the variance associated with $\mathcal{G}(u,\xi_i)$ and $\mathcal{H}(u,\zeta_i)$ is bounded. It should be noted that deterministic VIs, denoted by $VI(Z;G,H,J)$, are special cases of SVIs with $\sigma_G = \sigma_H = 0$. The above setting covers as a special case of the regular SVIs whose operators $G(u)$ or $H(u)$ are given in the form of expectation as shown in (1). Moreover, it provides a framework to study randomized algorithms for solving deterministic VI or saddle point problems [33] (see Sect. 4.2 for an example).

One popular approach based on Monte-carlo sampling to solve SVI is the sample average approximation (SAA) method [8,41,42,48,50]. In this approach one first approximates the expectation $\mathcal{F}(u)=\mathbb{E}[\mathcal{F}(u;\xi,\zeta)]$ by $\tilde{\mathcal{F}}(u)\equiv\sum_{i=1}^N\mathbb{E}[\mathcal{F}(u;\xi_i,\zeta_i)]/N$ for some $N>0$, and then solves a deterministic counterpart of (1) with $\mathcal{F}$ replaced by $\tilde{\mathcal{F}}$. However, the resulting deterministic VI problem is still often difficult to solve when the dimension of u or the sample size N is large. Moreover, this approach is not applicable to the online setting when the decision vector needs to be updated as new samples arrive. Recently, there has been a resurgence of interest in stochastic approximation (SA) type algorithms which aim at solving the SVIs directly based on the noisy estimation of the operators returned by the stochastic oracles (see, e.g., [19,20,23,24,33,51]). These more recent studies focused on analyzing the convergence behaviour of SA type methods during a finite number of iterations (i.e., complexity) and exploring whether these performance bounds are tight or not. However, to the best of our knowledge, none of existing algorithms can attain the theoretically optimal rate of convergence to solve the SVI problems in (1) due to its rich structural properties (e.g., gradient field G and Lipschitz continuity of H). More specifically, we can see that the total number of gradient and operator evaluations for solving SVI cannot be smaller than

$$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}+\frac{M}{\epsilon}+\frac{\sigma^2}{\epsilon^2}\right). \tag{9}$$

This is a lower complexity bound derived based on the following three observations:

1. If $H=0$ and $\sigma=0$, $SVI(Z;G,0,0)$ is equivalent to a smooth optimization problem $\min_{u\in Z}G(u)$, and the complexity for minimizing $G(u)$ cannot be better than $\mathcal{O}(\sqrt{L/\epsilon})$ [34,38];
2. If $G=0$ and $\sigma=0$, the complexity for solving $SVI(Z;0,H,0)$ cannot be better than $\mathcal{O}(M/\epsilon)$ [31] (see also the discussions in Section 5 of [32]).
3. If $H=0$, $SVI(Z;G,0,0)$ is equivalent to a stochastic smooth optimization problem, and the complexity cannot be better than $\mathcal{O}(\sigma^2/\epsilon^2)$ [24].

However, there exist significant gaps between the above lower complexity bound and the complexity of existing algorithms, especially in terms of their dependence on the Lipschitz constants L and M, while it is well-known that SA-type methods are sensitive to these parameters. It is worth noting that the above lower complexity bound has not been attained even for the deterministic VIs with $\sigma = 0$.

The lower complexity bound in (9) and the three observations stated above provide some important guidelines to the design of efficient algorithms to solve the SVI problem with the operator given in (5). It might seem natural to consider the more general problem (6) by combining $\nabla G(u)$ and $H(u)$ in (5) together as a single monotone operator, instead of separating them apart. Such consideration is reasonable from a generalization point of view, by noting that the convexity of function $G(u)$ is equivalent to the monotonicity of $\nabla G(u)$, and the Lipschitz conditions (3) and (4) are equivalent to a Lipschitz condition of $F(u)$ in (6) with $\|F(w) - F(v)\|_* \leq (L + M)\|w - v\|$. However, from the algorithmic point of view, a special treatment of ∇G separately from H is crucial for the design of accelerated algorithms. By observations 2 and 3 above, if we consider $F := \nabla G + H$ as a single monotone operator, the complexity for solving $SVI(Z; 0; F; 0)$ can not be smaller than

$$\mathcal{O}\left(\frac{L+M}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right).$$

This is worse than (9) in terms of the dependence on L. The identification and specialized treatment on the gradient field ∇G allows us to use its Lipschitz condition (4) to achieve the optimal iteration complexity (see observation 1 above) for solving SVIs.

In order to achieve the complexity bound in (9) for SVIs, we incorporate a multi-step acceleration scheme into the stochastic mirror-prox method in [20], and introduce a stochastic accelerated mirror-prox (SAMP) method that fully exploits the structural properties of (1). We show that SAMP can exhibit a complexity bound given by (9). To the best of our knowledge, this is the first time in the literature that the lower complexity bound in (9) has been achieved for SVIs. Table 1 shows in more details how our results improve the best-known so-far complexity results for solving deterministic and stochastic VIs. In particular, for deterministic VIs, the Lipschitz constant L can be as large as $\Omega(1/\epsilon)$ without affecting the rate of convergence. Moreover, the Lipschitz constant L can be as large as $\Omega(1/\epsilon^{3/2})$ without significantly slowing down the convergence rate for solving SVIs. We demonstrate the advantages of these accelerated algorithms over some existing algorithms through our numerical experiments in Sect. 4.

In addition to improving existing complexity bounds for solving VI problems, we incorporate into SAMP the termination criterion employed by Monteiro and Svaiter [28,29] for solving variational and hemivariational inequalities posed as monotone inclusion problem. As a result, the SAMP can deal with the case when Z is unbounded, as long as a strong solution to problem (6) exists, and the iteration complexity of SAMP will depend on the distance from the initial point to the set of strong solutions. It is worth noting that no such complexity results have been presented before in the literature for solving unbounded SVIs.

Table 1 Comparison of complexity results

Problem class	Complexity bound	Related work
$VI(Z; G, H, 0)$	$\mathcal{O}\left(\frac{L+M}{\varepsilon}\right)$	Nemirovski [32] (see also [1,35])
$VI(Z; G, H, J)$	$\mathcal{O}\left(\frac{L+M}{\varepsilon}\right)$	Monteiro et al. [29]
$VI(Z; G, H, J)$	$\mathcal{O}\left(\sqrt{\frac{L}{\varepsilon}}+\frac{M}{\varepsilon}\right)$	This paper
$SVI(Z; G, H, 0)$	$\mathcal{O}\left(\frac{L+M}{\varepsilon}+\frac{\sigma^2}{\varepsilon^2}\right)$	Juditsky et al. [20]
$SVI(Z; G, H, J)$	$\mathcal{O}\left(\sqrt{\frac{L}{\varepsilon}}+\frac{M}{\varepsilon}+\frac{\sigma^2}{\varepsilon^2}\right)$	This paper

The remaining part of this paper is organized as follows. We propose the SAMP algorithm and discuss its main convergence results for solving SVIs in Section 2. To facilitate the readers, we present the proofs of the main convergence results in Sect. 3. Some preliminary numerical experiments are provided in Sect. 4 to demonstrate the efficiency of the SAMP algorithm. Finally, we make some concluding remarks in Sect. 5.

2 The stochastic accelerated mirror-prox method

In this section, we develop the stochastic accelerated mirror-prox (SAMP) method for solving $SVI(Z; F)$ and demonstrate that it can achieve the optimal rate of convergence in (9).

Throughout this paper, we assume that the following *prox-mapping* can be solved efficiently:

$$P_z^J(\eta) := \operatorname*{argmin}_{u\in Z} \langle \eta, u - z\rangle + V(z, u) + J(u). \tag{10}$$

In (10), the function $V(\cdot,\cdot)$ is defined by

$$V(z, u) := \omega(u) - \omega(z) - \langle \nabla\omega(z), u - z\rangle, \quad \forall u,\ z \in Z, \tag{11}$$

where $\omega(\cdot)$ is a strongly convex function with convexity parameter $\mu > 0$, and is called the *distance generating function*. The function $V(\cdot,\cdot)$ is known as a *prox-function*, or *Bregman divergence* [6] (see, e.g., [2,3,32,39] for the properties of prox-functions and prox-mappings and their applications in convex optimization). Using the aforementioned definition of the prox-mapping, we describe the SAMP method in Algorithm 1.

Observe that in the SAMP algorithm we introduced two sequences, i.e., $\{w_t^{md}\}$ and $\{w_t^{ag}\}$ (here "md" stands for "middle", and "ag" stands for "aggregated"), that are convex combinations of iterations $\{w_t\}$ and $\{r_t\}$ as long as $\alpha_t \in [0, 1]$. If $\alpha_t \equiv 1$,

Algorithm 1 The stochastic accelerated mirror-prox (SAMP) method

Choose $r_1 \in Z$. Set $w_1 = r_1$, $w_1^{ag} = r_1$.
For $t = 1, 2, \ldots, N-1$, calculate

$$w_t^{md} = (1-\alpha_t) w_t^{ag} + \alpha_t r_t, \tag{12}$$

$$w_{t+1} = P_{r_t}^{\gamma_t J}\left(\gamma_t \mathcal{H}(r_t; \zeta_{2t-1}) + \gamma_t \mathcal{G}(w_t^{md}; \xi_t)\right), \tag{13}$$

$$r_{t+1} = P_{r_t}^{\gamma_t J}\left(\gamma_t \mathcal{H}(w_{t+1}; \zeta_{2t}) + \gamma_t \mathcal{G}(w_t^{md}; \xi_t)\right), \tag{14}$$

$$w_{t+1}^{ag} = (1-\alpha_t) w_t^{ag} + \alpha_t w_{t+1}. \tag{15}$$

Output w_N^{ag}.

$G = 0$ and $J = 0$, then Algorithm 1 for solving $SVI(Z; 0, H, 0)$ is equivalent to the stochastic mirror-prox method in [20]. If, in addition, $\sigma = 0$, then it reduces to the mirror-prox method in [32]. Moreover, if the distance generating function $w(\cdot) = \|\cdot\|^2/2$, then iterations (13) and (14) become

$$w_{t+1} = \operatorname*{argmin}_{u \in Z} \langle \gamma_t H(r_t), u - r_t \rangle + \frac{1}{2}\|u - r_t\|^2,$$
$$r_{t+1} = \operatorname*{argmin}_{u \in Z} \langle \gamma_t H(w_{t+1}), u - r_t \rangle + \frac{1}{2}\|u - r_t\|^2,$$

which are exactly the iterates of the extragradient method in [22]. On the other hand, if $H = 0$, then (13) and (14) produce the same optimizer $w_{t+1} = r_{t+1}$, and Algorithm 1 is equivalent to the accelerated stochastic approximation method in [24]. Specifically, if, in addition, $\sigma = 0$, then it reduces to a version of Nesterov's accelerated gradient method for solving $\min_{u \in Z} G(u) + J(u)$ (see, for example, Algorithm 1 in [46]). Therefore, Algorithm 1 can be viewed as a hybrid algorithm of the stochastic mirror-prox method and the accelerated stochastic approximation method, which gives its name stochastic accelerated mirror-prox method. It is interesting to note that for any t, there are two calls of $\mathcal{SO}_H$ but just one call of $\mathcal{SO}_G$. However, if we assume that $J = 0$ and use the stochastic mirror-prox method in [20] to solve $SVI(Z; G, H, 0)$, for any t there would be two calls of $\mathcal{SO}_H$ and two calls of $\mathcal{SO}_G$. Therefore, the cost per iteration of SAMP is less than that of the stochastic mirror-prox method.

In order to analyze the convergence of Algorithm 1, we introduce a notion to characterize the weak solutions of $SVI(Z; G, H, J)$. For all $\tilde{u}, u \in Z$, we define

$$Q(\tilde{u}, u) := G(\tilde{u}) - G(u) + \langle H(u), \tilde{u} - u \rangle + J(\tilde{u}) - J(u). \tag{16}$$

Clearly, for F defined in (5), we have $\langle F(u), \tilde{u} - u\rangle \leq Q(\tilde{u}, u)$. Therefore, if $Q(\tilde{u}, u) \leq 0$ for all $u \in Z$, then $\tilde{u}$ is a weak solution of $SVI(Z; G, H, J)$. Hence when Z is bounded, it is natural to use the gap function

$$g(\tilde{u}) := \sup_{u \in Z} Q(\tilde{u}, u) \tag{17}$$

to evaluate the accuracy of a feasible solution $\tilde{u} \in Z$. However, if Z is unbounded, then $g(\tilde{z})$ may not be well-defined, even when $\tilde{z} \in Z$ is a nearly optimal solution. Therefore, we need to employ a slightly modified gap function in order to measure the accuracy of candidate solutions when Z is unbounded. In the sequel, we will consider the cases of bounded and unbounded Z separately. For both cases we establish the rate of convergence of the gap functions in terms of their expectation, i.e., the "average" rate of convergence over many runs of the algorithm. Furthermore, we demonstrate that if Z is bounded, then we can also establish the rate of convergence of $g(\cdot)$ in the probability sense, under the following "light-tail" assumption:

A2 For any i-th call on oracles $\mathcal{SO}_H$ and $\mathcal{SO}_H$ with any input $u \in Z$,

$$\mathbb{E}[\exp\{\|\nabla G(u) - \mathcal{G}(u; \xi_i)\|_*^2 / \sigma_G^2\}] \leq \exp\{1\},$$

and

$$\mathbb{E}[\exp\{\|H(u) - \mathcal{H}(u; \zeta_i)\|_*^2 / \sigma_H^2\}] \leq \exp\{1\}.$$

Assumption A2 is sometimes called the sub-Gaussian assumption. Many different random variables, such as Gaussian, uniform, and any random variables with a bounded support, will satisfy this assumption. It should be noted that Assumption A2 implies Assumption A1 by Jensen's inequality.

We start with establishing some convergence properties of Algorithm 1 when Z is bounded. It should be noted that the following quantity will be used throughout the convergence analysis of this paper:

$$\Gamma_t = \begin{cases} 1, & \text{when } t = 1 \\ (1 - \alpha_t)\Gamma_{t-1}, & \text{when } t > 1. \end{cases} \tag{18}$$

Theorem 1 *Suppose that*

$$\sup_{z_1, z_2 \in Z} V(z_1, z_2) \leq \Omega_Z^2. \tag{19}$$

Also assume that the parameters $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *in Algorithm 1 satisfy* $\alpha_1 = 1$,

$$q\mu - L\alpha_t\gamma_t - \frac{3M^2\gamma_t^2}{\mu} \geq 0 \quad \text{for some } q \in (0, 1), \text{ and } \frac{\alpha_t}{\Gamma_t\gamma_t} \leq \frac{\alpha_{t+1}}{\Gamma_{t+1}\gamma_{t+1}}, \quad \forall t \geq 1, \tag{20}$$

where Γ_t *is defined in* (18). *Then,*

(a) Under Assumption A1, for all $t \geq 1$,

$$\mathbb{E}\left[g(w_{t+1}^{ag})\right] \leq \mathcal{Q}_0(t) := \frac{2\alpha_t}{\gamma_t}\Omega_Z^2 + \left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right]\Gamma_t \sum_{i=1}^{t} \frac{\alpha_i \gamma_i}{\mu \Gamma_i}. \tag{21}$$

(b) Under Assumption A2, for all $\lambda > 0$ and $t \geq 1$,

$$Prob\{g(w_{t+1}^{ag}) > \mathcal{Q}_0(t) + \lambda \mathcal{Q}_1(t)\} \leq 2\exp\{-\lambda^2/3\} + 3\exp\{-\lambda\}, \tag{22}$$

where

$$\begin{aligned} \mathcal{Q}_1(t) := \Gamma_t(\sigma_G + \sigma_H)\Omega_Z \sqrt{\frac{2}{\mu}\sum_{i=1}^{t}\left(\frac{\alpha_i}{\Gamma_i}\right)^2} \\ + \left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right]\Gamma_t \sum_{i=1}^{t} \frac{\alpha_i \gamma_i}{\mu \Gamma_i}. \end{aligned} \tag{23}$$

There are various options for choosing the parameters $\{\alpha_t\}$ and $\{\gamma_t\}$ that satisfy (20). In the following corollary, we give one example of such parameter settings.

Corollary 1 *Suppose that (19) holds. If the stepsizes $\{\alpha_t\}$ and $\{\gamma_t\}$ in Algorithm 1 are set to:*

$$\alpha_t = \frac{2}{t+1} \text{ and } \gamma_t = \frac{\mu t}{4L + 3Mt + \beta(t+1)\sqrt{\mu t}}, \tag{24}$$

where $\beta > 0$ is a parameter. Then under Assumption A1,

$$\begin{aligned} \mathbb{E}\left[g(w_{t+1}^{ag})\right] \leq \frac{16L\Omega_Z^2}{\mu t(t+1)} + \frac{12M\Omega_Z^2}{\mu(t+1)} \\ + \frac{\sigma\Omega_Z}{\sqrt{\mu(t-1)}}\left(\frac{4\beta\Omega_Z}{\sigma} + \frac{16\sigma}{3\beta\Omega_Z}\right) =: \mathcal{C}_0(t), \end{aligned} \tag{25}$$

where σ and Ω_Z are defined in (8) and (19), respectively. Furthermore, under Assumption A2,

$$Prob\{g(w_{t+1}^{ag}) > \mathcal{C}_0(t) + \lambda \mathcal{C}_1(t)\} \leq 2\exp\{-\lambda^2/3\} + 3\exp\{-\lambda\}, \quad \forall \lambda > 0,$$

where

$$\mathcal{C}_1(t) := \frac{\sigma\Omega_Z}{\sqrt{\mu(t-1)}}\left(\frac{4\sqrt{3}}{3} + \frac{16\sigma}{3\beta\Omega_Z}\right). \tag{26}$$

Proof It is easy to check that

$$\Gamma_t = \frac{2}{t(t+1)} \text{ and } \frac{\alpha_t}{\Gamma_t \gamma_t} \le \frac{\alpha_{t+1}}{\Gamma_{t+1}\gamma_{t+1}}.$$

In addition, in view of (24), we have $\gamma_t \le \mu t/(4L)$ and $\gamma_t^2 \le (\mu^2)/(9M^2)$, which implies

$$\frac{5\mu}{6} - L\alpha_t\gamma_t - \frac{3M^2\gamma_t^2}{\mu} \ge \frac{5\mu}{6} - \frac{\mu t}{4}\cdot\frac{2}{t+1} - \frac{\mu}{3} \ge 0.$$

Therefore the first relation in (20) holds with constant $q = 5/6$. In view of Theorem 1, it now suffices to show that $\mathcal{Q}_0(t) \le \mathcal{C}_0(t)$ and $\mathcal{Q}_1(t) \le \mathcal{C}_1(t)$. Observing that $\alpha_t/\Gamma_t = t$, and $\gamma_t \le \sqrt{\mu}/(\beta\sqrt{t})$, we obtain

$$\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\Gamma_i} \le \frac{\sqrt{\mu}}{\beta}\sum_{i=1}^{t}\sqrt{i} \le \frac{\sqrt{\mu}}{\beta}\int_0^{t+1}\sqrt{t}dt = \frac{\sqrt{\mu}}{\beta}\cdot\frac{2(t+1)^{3/2}}{3} = \frac{2\sqrt{\mu}(t+1)^{3/2}}{3\beta}.$$

Using the above relation, (19), (21), (23), (24), and the fact that $\sqrt{t+1}/t \le 1/\sqrt{t-1}$ and $\sum_{i=1}^{t} i^2 \le t(t+1)^2/3$, we have

$$\begin{aligned}\mathcal{Q}_0(t) &= \frac{4\Omega_Z^2}{\mu t(t+1)}\left(4L + 3Mt + \beta(t+1)\sqrt{\mu t}\right) + \frac{8\sigma^2}{\mu t(t+1)}\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\Gamma_i}\\ &\le \frac{16L\Omega_Z^2}{\mu t(t+1)} + \frac{12M\Omega_Z^2}{\mu(t+1)} + \frac{4\beta\Omega_Z^2}{\sqrt{\mu t}} + \frac{16\sigma^2\sqrt{t+1}}{3\sqrt{\mu}\beta t}\\ &\le \mathcal{C}_0(t),\end{aligned}$$

and

$$\begin{aligned}\mathcal{Q}_1(t) &= \frac{2(\sigma_G+\sigma_H)}{t(t+1)}\Omega_Z\sqrt{\frac{2}{\mu}\sum_{i=1}^{t}i^2} + \frac{8\sigma^2}{\mu t(t+1)}\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\Gamma_i}\\ &\le \frac{2\sqrt{2}(\sigma_G+\sigma_H)\Omega_Z}{\sqrt{3\mu t}} + \frac{16\sigma^2\sqrt{t+1}}{3\sqrt{\mu}\beta t}\\ &\le \mathcal{C}_1(t).\end{aligned}$$

We now add a few remarks about the results obtained in Corollary 1. Firstly, in view of (9), (25) and (26), we can clearly see that the SAMP method is robust with respect to the estimates of σ and Ω_Z. Indeed, the SAMP method achieves the optimal iteration complexity for solving the SVI problem as long as $\beta = \mathcal{O}(\sigma/\Omega_Z)$. In particular, in this case, the number of iterations performed by the AMP method to find an ϵ-solution of (1), i.e., a point $\bar{w} \in Z$ s.t. $\mathbb{E}[g(\bar{w})] \le \epsilon$, can be bounded by

$$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}+\frac{M}{\epsilon}+\frac{\sigma^2}{\epsilon^2}\right), \tag{27}$$

which implies that this algorithm allows L to be as large as $\mathcal{O}(\epsilon^{-3/2})$ and M to be as large as $\mathcal{O}(\epsilon^{-1})$ without significantly affecting its convergence properties. Secondly, for the deterministic case when $\sigma = 0$, the complexity bound in (27) significantly improves the best-known so-far complexity for solving problem (1) (see (9)) in terms of their dependence on the Lipschitz constant L.

In the following theorem, we demonstrate some convergence properties of Algorithm 1 for solving the stochastic problem $SVI(Z; G, H, J)$ when Z is unbounded. It seems that this case has not been well-studied previously in the literature. To study the convergence properties of AMP in this case, we use a perturbation-based termination criterion recently employed by Monteiro and Svaiter [28,29], which is based on the enlargement of a maximal monotone operator first introduced in [7]. More specifically, we say that the pair $(\tilde{v}, \tilde{u}) \in \mathcal{E} \times Z$ is a (ρ, ε)-approximate solution of $SVI(Z; G, H, J)$ if $\|\tilde{v}\| \leq \rho$ and $\tilde{g}(\tilde{u}, \tilde{v}) \leq \varepsilon$, where the gap function $\tilde{g}(\cdot, \cdot)$ is defined by

$$\tilde{g}(\tilde{u}, \tilde{v}) := \sup_{u \in Z} Q(\tilde{u}, u) - \langle \tilde{v}, \tilde{u} - u \rangle. \tag{28}$$

We call $\tilde{v}$ the perturbation vector associated with $\tilde{u}$. One advantage of employing this termination criterion is that the convergence analysis does not depend on the boundedness of Z.

Theorem 2 below describes the convergence properties of SAMP for solving SVIs with unbounded feasible sets, under the assumption that a strong solution of (6) exists. It should be noted that this assumption does not limit too much the applicability of the SAMP method. For example, when $J \equiv 0$ in (1), the conditions for the existence of strong solutions are described in Section 2.2 of the seminal book [12]. Indeed, any weak solution to $SVI(Z; F)$ is also a strong solution in such case. For general case in which $J \not\equiv 0$ and F in (5) is a point-to-set map, there has also been several studies on the theories of strong solutions to (6) (see, e.g., [13,21]). For example, when J is a finite-valued closed convex function, some conditions for the existence of strong solutions are proven in Theorem 2.3 of [21].

Theorem 2 *Suppose that $V(r, z) := \|z - r\|^2/2$ for any $r \in Z$ and $z \in Z$. If the parameters $\{\alpha_t\}$ and $\{\gamma_t\}$ in Algorithm 1 are chosen such that $\alpha_1 = 1$, and for all $t > 1$,*

$$0 \leq \alpha_t < 1,\ L\alpha_t\gamma_t + 3M^2\gamma_t^2 \leq c^2 < q \text{ for some } c, q \in (0, 1),\ \text{and } \frac{\alpha_t}{\Gamma_t\gamma_t} = \frac{\alpha_{t+1}}{\Gamma_{t+1}\gamma_{t+1}}, \tag{29}$$

where Γ_t is defined in (18). Then for all $t \geq 1$ there exists a perturbation vector v_{t+1} and a residual $\varepsilon_{t+1} \geq 0$ such that $\tilde{g}(w_{t+1}^{ag}, v_{t+1}) \leq \varepsilon_{t+1}$. Moreover, for all $t \geq 1$, we have

$$\mathbb{E}[\|v_{t+1}\|] \leq \frac{\alpha_t}{\gamma_t}\left(2D + 2\sqrt{D^2 + C_t^2}\right), \tag{30}$$

$$\mathbb{E}[\varepsilon_{t+1}] \leq \frac{\alpha_t}{\gamma_t}\left[(3+6\theta)D^2 + (1+6\theta)C_t^2\right] + \frac{18\alpha_t^2\sigma_H^2}{\gamma_t^2}\sum_{i=1}^{t}\gamma_i^3, \tag{31}$$

where

$$D := \|r_1 - u^*\|, \tag{32}$$

u^* *is a strong solution of* $SVI(Z; G, H, J)$,

$$\theta = \max\left\{1, \frac{c^2}{q-c^2}\right\} \text{ and } C_t = \sqrt{\left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right]\sum_{i=1}^{t}\gamma_i^2}. \tag{33}$$

Below we give an example of parameters α_t and γ_t that satisfies (29).

Corollary 2 *Suppose that there exists a strong solution of* (1). *If the maximum number of iterations* N *is given, and the stepsizes* $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *in Algorithm* 1 *are set to*

$$\alpha_t = \frac{2}{t+1} \text{ and } \gamma_t = \frac{t}{5L + 3MN + \beta N\sqrt{N-1}}, \tag{34}$$

where σ *is defined in Corollary* 1, *then there exists* $v_N \in \mathcal{E}$ *and* $\varepsilon_N > 0$, *such that* $\tilde{g}(w_t^{ag}[][N], v_N) \leq \varepsilon_N$,

$$\mathbb{E}[\|v_N\|] \leq \frac{40LD}{N(N-1)} + \frac{24MD}{N-1} + \frac{\sigma}{\sqrt{N-1}}\left(\frac{8\beta D}{\sigma} + 5\right), \tag{35}$$

and

$$\mathbb{E}[\varepsilon_N] \leq \frac{90LD^2}{N(N-1)} + \frac{54MD^2}{N-1} + \frac{\sigma D}{\sqrt{N-1}}\left(\frac{18\beta D}{\sigma} + \frac{56\sigma}{3\beta D} + \frac{18\sigma}{\beta DN}\right). \tag{36}$$

Proof Clearly, we have $\Gamma_t = 2/[t(t+1)]$, and hence (18) is satisfied. Moreover, in view of (34), we have

$$\begin{aligned} L\alpha_t\gamma_t + 3M^2\gamma_t^2 &\leq \frac{2L}{5L+3MN} + \frac{3M^2N^2}{(5L+3MN)^2} \\ &= \frac{10L^2 + 6LMN + 3M^2N^2}{(5L+3MN)^2} < \frac{5}{12} < \frac{5}{6}, \end{aligned}$$

which implies that (29) is satisfied with $c^2 = 5/12$ and $q = 5/6$. Observing from (34) that $\gamma_t = t\gamma_1$, setting $t = N-1$ in (33) and (34), we obtain

$$\frac{\alpha_{N-1}}{\gamma_{N-1}} = \frac{2}{\gamma_1 N(N-1)} \text{ and } C_{N-1}^2 = 4\sigma^2\sum_{i=1}^{N-1}\gamma_1^2 i^2 \leq \frac{4\sigma^2\gamma_1^2N^2(N-1)}{3}, \tag{37}$$

where C_{N-1} is defined in (33). Applying (37) to (30) we have

$$\begin{aligned}\mathbb{E}[\|v_N\|] &\le \frac{2}{\gamma_1 N(N-1)}(4D + 2C_{N-1}) \le \frac{8D}{\gamma_1 N(N-1)} + \frac{8\sigma}{\sqrt{3(N-1)}} \\ &\le \frac{40LD}{N(N-1)} + \frac{24MD}{N-1} + \frac{\sigma}{\sqrt{N-1}}\left(\frac{8\beta D}{\sigma} + 5\right).\end{aligned}$$

In addition, using (31), (37), and the facts that $\theta = 1$ in (33) and

$$\sum_{i=1}^{N-1} \gamma_i^3 = \gamma_1^3 N^2 (N-1)^2/4,$$

we have

$$\begin{aligned}\mathbb{E}[\varepsilon_{N-1}] &\le \frac{2}{\gamma_1 N(N-1)}(9D^2 + 7C_{N-1}^2) + \frac{72\sigma_H^2}{\gamma_1^2 N^2 (N-1)^2} \cdot \frac{\gamma_1^3 N^2 (N-1)^2}{4} \\ &\le \frac{18D^2}{\gamma_1 N(N-1)} + \frac{56\sigma^2 \gamma_1 N}{3} + 18\sigma_H^2 \gamma_1 \\ &\le \frac{90LD^2}{N(N-1)} + \frac{54MD^2}{N-1} + \frac{18\beta D^2}{\sqrt{N-1}} + \frac{56\sigma^2}{3\beta\sqrt{N-1}} + \frac{18\sigma_H^2}{\beta N\sqrt{N-1}} \\ &\le \frac{90LD^2}{N(N-1)} + \frac{54MD^2}{N-1} + \frac{\sigma D}{\sqrt{N-1}}\left(\frac{18\beta D}{\sigma} + \frac{56\sigma}{3\beta D} + \frac{18\sigma}{\beta DN}\right).\end{aligned}$$

Several remarks are in place for the results obtained in Theorem 2 and Corollary 2. Firstly, similarly to the bounded case (see the remark after Corollary 1), one may want to choose β in a way such that the right hand side of (35) or (36) is minimized, e.g., $\beta = \mathcal{O}(\sigma/D)$. However, since the value of D will be very difficult to estimate for the unbounded case and hence one often has to resort to a suboptimal selection for β. For example, if $\beta = \sigma$, then the RHS of (35) and (36) will become $\mathcal{O}(LD/N^2 + MD/N + \sigma D/\sqrt{N})$ and $\mathcal{O}(LD^2/N^2 + MD^2/N + \sigma D^2/\sqrt{N})$, respectively. Secondly, both residuals $\|v_N\|$ and ε_N in (35) and (36) converge to 0 at the same rate (up to a constant factor). Finally, it is only for simplicity that we assume that $V(r, z) = \|z - r\|^2/2$; Similar results can be achieved under assumptions that $\nabla\omega$ is Lipschitz continuous.

3 Convergence analysis

In this section, we focus on proving the main convergence results in Sect. 2, namely, Theorems 1 and 2.

To prove the convergence of the stochastic AMP algorithm, we first present some technical results. Lemmas 1 and 2 describe some important properties of the prox-mapping $P_r^J(\eta)$ used in (13) and (14) of Algorithm 1. Lemma 3 provides a recursion related to the function $Q(\cdot, \cdot)$ defined in (16). With the help of Lemmas 1, 2 and 3, we estimate a bound on $Q(\cdot, \cdot)$ in Lemma 4.

Lemma 1 *For all $r, \zeta \in \mathcal{E}$, if $w = P_r^J(\zeta)$, then for all $u \in Z$, we have*

$$\langle \zeta, w - u \rangle + J(w) - J(u) \leq V(r, u) - V(r, w) - V(w, u).$$

Proof See Lemma 2 in [14] for the proof.

The following lemma is a slight extension of Lemma 6.3 in [20]. In particular, when $J(\cdot) = 0$, we can obtain (41) and (42) directly by applying (40) to (6.8) in [20], and the results when $J(\cdot) \not\equiv 0$ can be easily constructed from the proof of Lemma 6.3 in [20]. We provide the proof here only for the sake of completeness.

Lemma 2 *Given $r, w, y \in Z$ and $\eta, \vartheta \in \mathcal{E}$ that satisfy*

$$w = P_r^J(\eta), \tag{38}$$
$$y = P_r^J(\vartheta), \tag{39}$$

and

$$\|\vartheta - \eta\|_*^2 \leq L^2 \|w - r\|^2 + M^2. \tag{40}$$

Then, for all $u \in Z$,

$$\langle \vartheta, w - u \rangle + J(w) - J(u) \leq V(r, u) - V(y, u) - \left(\frac{\mu}{2} - \frac{L^2}{2\mu} \right) \|r - w\|^2 + \frac{M^2}{2\mu}, \tag{41}$$

and

$$V(y, w) \leq \frac{L^2}{\mu^2} V(r, w) + \frac{M^2}{2\mu}. \tag{42}$$

Proof Applying Lemma 1 to (38) and (39), for all $u \in Z$ we have

$$\langle \eta, w - u \rangle + J(w) - J(u) \leq V(r, u) - V(r, w) - V(w, u), \tag{43}$$
$$\langle \vartheta, y - u \rangle + J(y) - J(u) \leq V(r, u) - V(r, y) - V(y, u), \tag{44}$$

In particular, letting $u = y$ in (43) we have

$$\langle \eta, w - y \rangle + J(w) - J(y) \leq V(r, y) - V(r, w) - V(w, y). \tag{45}$$

Adding inequalities (44) and (45), then

$$\begin{aligned} &\langle \vartheta, y - u \rangle + \langle \eta, w - y \rangle + J(w) - J(u) \\ &\quad \leq V(r, u) - V(y, u) - V(r, w) - V(w, y), \end{aligned}$$

which is equivalent to

$$\langle \vartheta, w-u\rangle + J(w) - J(u) \le \langle \vartheta-\eta, w-y\rangle + V(r,u) - V(y,u) - V(r,w) - V(w,y).$$

Applying Schwartz inequality and Young's inequality to the above inequality, and using the fact that

$$\frac{\mu}{2}\|z-u\|^2 \le V(u,z), \forall u, z \in Z, \tag{46}$$

due to the strong convexity of $\omega(\cdot)$ in (11), we obtain

$$\begin{aligned}
&\langle \vartheta, w-u\rangle + J(w) - J(u)\\
&\le \|\vartheta-\eta\|_*\|w-y\| + V(r,u) - V(y,u) - V(r,w) - \frac{\mu}{2}\|w-y\|^2\\
&\le \frac{1}{2\mu}\|\vartheta-\eta\|_*^2 + \frac{\mu}{2}\|w-y\|^2 + V(r,u) - V(y,u) - V(r,w) - \frac{\mu}{2}\|w-y\|^2\\
&= \frac{1}{2\mu}\|\vartheta-\eta\|_*^2 + V(r,u) - V(y,u) - V(r,w).
\end{aligned} \tag{47}$$

The result in (41) then follows immediately from above relation, (40) and (46).

Moreover, observe that by setting $u = w$ and $u = y$ in (44) and (47), respectively, we have

$$\begin{aligned}
\langle \vartheta, y-w\rangle + J(y) - J(w) &\le V(r,w) - V(r,y) - V(y,w),\\
\langle \vartheta, w-y\rangle + J(w) - J(y) &\le \frac{1}{2\mu}\|\vartheta-\eta\|_*^2 + V(r,y) - V(r,w).
\end{aligned}$$

Adding the above two inequalities, and using (40) and (46), we have

$$\begin{aligned}
0 &\le \frac{1}{2\mu}\|\vartheta-\eta\|_*^2 - V(y,w) \le \frac{L^2}{2\mu}\|r-w\|^2 + \frac{M^2}{2\mu} - V(y,w)\\
&\le \frac{L^2}{\mu^2}V(r,w) + \frac{M^2}{2\mu} - V(y,w),
\end{aligned}$$

and thus (42) holds.

Lemma 3 *For any sequences $\{r_t\}_{t\ge1}$ and $\{w_t\}_{t\ge1} \subset Z$, if the sequences $\{w_t^{ag}\}$ and $\{w_t^{md}\}$ are generated by (12) and (15), then for all $u \in Z$,*

$$\begin{aligned}
Q(w_{t+1}^{ag}, u) - (1-\alpha_t)Q(w_t^{ag}, u) &\le \alpha_t\langle \nabla G(w_t^{md}) + H(w_{t+1}), w_{t+1}-u\rangle\\
&+ \frac{L\alpha_t^2}{2}\|w_{t+1}-r_t\|^2 + \alpha_t J(w_{t+1}) - \alpha_t J(u).
\end{aligned} \tag{48}$$

Proof Observe from (12) and (15) that $w_{t+1}^{ag} - w_t^{md} = \alpha_t(w_{t+1} - r_t)$. This observation together with the convexity of $G(\cdot)$ imply that for all $u \in Z$,

$$\begin{aligned} G(w_{t+1}^{ag}) \le\ & G(w_t^{md}) + \langle \nabla G(w_t^{md}), w_{t+1}^{ag} - w_t^{md} \rangle + \frac{L}{2}\|w_{t+1}^{ag} - w_t^{md}\|^2 \\ =\ & (1-\alpha_t)\left[G(w_t^{md}) + \langle \nabla G(w_t^{md}), w_t^{ag} - w_t^{md}\rangle\right] \\ & + \alpha_t\left[G(w_t^{md}) + \langle \nabla G(w_t^{md}), u - w_t^{md}\rangle\right] \\ & + \alpha_t\langle \nabla G(w_t^{md}), w_{t+1} - u\rangle + \frac{L\alpha_t^2}{2}\|w_{t+1} - r_t\|^2 \\ \le\ & (1-\alpha_t)G(w_t^{ag}) + \alpha_t G(u) + \alpha_t\langle \nabla G(w_t^{md}), w_{t+1} - u\rangle \\ & + \frac{L\alpha_t^2}{2}\|w_{t+1} - r_t\|^2. \end{aligned}$$

Using the above inequality, (15), (16) and the monotonicity of $H(\cdot)$, we have

$$\begin{aligned} & Q(w_{t+1}^{ag}, u) - (1-\alpha_t)Q(w_t^{ag}, u) \\ =\ & G(w_{t+1}^{ag}) - (1-\alpha_t)G(w_t^{ag}) - \alpha_t G(u) \\ & + \langle H(u), w_{t+1}^{ag} - u\rangle - (1-\alpha_t)\langle H(u), w_t^{ag} - u\rangle \\ & + J(w_{t+1}^{ag}) - (1-\alpha_t)J(w_t^{ag}) - \alpha_t J(u) \\ \le\ & G(w_{t+1}^{ag}) - (1-\alpha_t)G(w_t^{ag}) - \alpha_t G(u) + \alpha_t\langle H(u), w_{t+1} - u\rangle \\ & + \alpha_t J(w_{t+1}) - \alpha_t J(u) \\ \le\ & \alpha_t\langle \nabla G(w_t^{md}), w_{t+1} - u\rangle + \frac{L\alpha_t^2}{2}\|w_{t+1} - r_t\|^2 + \alpha_t\langle H(w_{t+1}), w_{t+1} - u\rangle \\ & + \alpha_t J(w_{t+1}) - \alpha_t J(u). \end{aligned}$$

In the sequel, we will use the following notations to describe the inexactness of the first order information from $\mathcal{SO}_H$ and $\mathcal{SO}_G$. At the t-th iteration, letting $\mathcal{H}(r_t; \zeta_{2t-1})$, $\mathcal{H}(w_{t+1}; \zeta_{2t})$ and $\mathcal{G}(w_t^{md}; \xi_t)$ be the output of the stochastic oracles, we denote

$$\begin{aligned} \Delta_H^{2t-1} &:= \mathcal{H}(r_t; \zeta_{2t-1}) - H(r_t), \\ \Delta_H^{2t} &:= \mathcal{H}(w_{t+1}; \zeta_{2t}) - H(w_{t+1}), \text{ and} \\ \Delta_G^t &:= \mathcal{G}(w_t^{md}; \xi_t) - \nabla G(w_t^{md}). \end{aligned} \tag{49}$$

Lemma 4 below provides a bound on $Q(w_{t+1}^{ag}, u)$ for all $u \in Z$.

Lemma 4 *Suppose that the parameters* $\{\alpha_t\}$ *in Algorithm 1 satisfies* $\alpha_1 = 1$ *and* $0 \le \alpha_t < 1$ *for all* $t > 1$. *Then the iterates* $\{r_t\}$, $\{w_t\}$ *and* $\{w_t^{ag}\}$ *satisfy*

$$\frac{1}{\Gamma_t} Q(w_{t+1}^{ag}, u) \le \mathcal{B}_t(u, r_{[t]}) - \sum_{i=1}^{t} \frac{\alpha_i}{2\Gamma_i \gamma_i} \left(q\mu - L\alpha_i \gamma_i - \frac{3M^2 \gamma_i^2}{\mu} \right) \|r_i - w_{i+1}\|^2 + \sum_{i=1}^{t} \Lambda_i(u), \ \forall u \in Z, \tag{50}$$

where Γ_t *is defined in* (18),

$$\mathcal{B}_t(u, r_{[t]}) := \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i \gamma_i} (V(r_i, u) - V(r_{i+1}, u)), \tag{51}$$

and

$$\Lambda_i(u) := \frac{3\alpha_i \gamma_i}{2\mu \Gamma_i} \left(\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2 \right) - \frac{(1-q)\mu \alpha_i}{2\Gamma_i \gamma_i} \|r_i - w_{i+1}\|^2 - \frac{\alpha_i}{\Gamma_i} \langle \Delta_H^{2i} + \Delta_G^i, w_{i+1} - u \rangle. \tag{52}$$

Proof Observe from (49) that

$$\begin{aligned} &\|\mathcal{H}(w_{t+1}; \zeta_{2t}) - \mathcal{H}(r_t; \zeta_{2t-1})\|_*^2 \\ &\le \left(\|H(w_{t+1}) - H(r_t)\|_* + \|\Delta_H^{2t}\|_* + \|\Delta_H^{2t-1}\|_* \right)^2 \\ &\le 3 \left(\|H(w_{t+1}) - H(r_t)\|_*^2 + \|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2 \right) \\ &\le 3 \left(M^2 \|w_{t+1} - r_t\|^2 + \|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2 \right). \end{aligned} \tag{53}$$

Applying Lemma 2 to (13) and (14) (with $r = r_t$, $w = w_{t+1}$, $y = r_{t+1}$, $\eta = \gamma_t \mathcal{H}(r_t; \zeta_{2t-1}) + \gamma_t \mathcal{G}(w_t^{md}; \xi_t)$, $\vartheta = \gamma_t \mathcal{H}(w_{t+1}; \zeta_{2t}) + \gamma_t \mathcal{G}(w_t^{md}; \xi_t)$, $J = \gamma_t J$, $L^2 = 3M^2\gamma_t^2$ and $M^2 = 3\gamma_t^2(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2)$), and using (53), we have for any $u \in Z$,

$$\begin{aligned} &\gamma_t \langle \mathcal{H}(w_{t+1}; \zeta_{2t}) + \mathcal{G}(w_t^{md}; \xi_t), w_{t+1} - u \rangle + \gamma_t J(w_{t+1}) - \gamma_t J(u) \\ &\le V(r_t, u) - V(r_{t+1}, u) - \left(\frac{\mu}{2} - \frac{3M^2\gamma_t^2}{2\mu} \right) \|r_t - w_{t+1}\|^2 \\ &\quad + \frac{3\gamma_t^2}{2\mu} (\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2). \end{aligned}$$

Applying (49) and the above inequality to (48), we have

$$\begin{aligned}
&Q(w_{t+1}^{ag}, u) - (1-\alpha_t)Q(w_t^{ag}, u) \\
&\le \alpha_t \langle \mathcal{H}(w_{t+1}; \zeta_{2t}) + \mathcal{G}(w_t^{md}; \xi_t), w_{t+1} - u\rangle + \alpha_t J(w_{t+1}) - \alpha_t J(u) \\
&\quad + \frac{L\alpha_t^2}{2}\|w_{t+1} - r_t\|^2 - \alpha_t \langle \Delta_H^{2t} + \Delta_G^t, w_{t+1} - u\rangle \\
&\le \frac{\alpha_t}{\gamma_t}\left(V(r_t, u) - V(r_{t+1}, u)\right) - \frac{\alpha_t}{2\gamma_t}\left(\mu - L\alpha_t\gamma_t - \frac{3M^2\gamma_t^2}{\mu}\right)\|r_t - w_{t+1}\|^2 \\
&\quad + \frac{3\alpha_t\gamma_t}{2\mu}\left(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2\right) - \alpha_t\langle \Delta_H^{2t} + \Delta_G^t, w_{t+1} - u\rangle.
\end{aligned}$$

Dividing the above inequality by Γ_t and using the definition of $\Lambda_t(u)$ in (52), we obtain

$$\begin{aligned}
&\frac{1}{\Gamma_t}Q(w_{t+1}^{ag}, u) - \frac{1-\alpha_t}{\Gamma_t}Q(w_t^{ag}, u) \\
&\le \frac{\alpha_t}{\Gamma_t\gamma_t}\left(V(r_t, u) - V(r_{t+1}, u)\right) \\
&\quad - \frac{\alpha_t}{2\Gamma_t\gamma_t}\left(q\mu - L\alpha_t\gamma_t - \frac{3M^2\gamma_t^2}{\mu}\right)\|r_t - w_{t+1}\|^2 + \Lambda_t(u).
\end{aligned}$$

Noting the fact that $\alpha_1 = 1$ and $(1-\alpha_t)/\Gamma_t = 1/\Gamma_{t-1}, t > 1$, due to (18), applying the above inequality recursively and using the definition of $\mathcal{B}_t(\cdot,\cdot)$ in (51), we conclude (50).

We still need the following technical result that helps to provide a bound on the last stochastic term in (50) before proving Theorems 1 and 2.

Lemma 5 *Let $\theta_t, \gamma_t > 0$, $t = 1, 2, \ldots$, be given. For any $w_1 \in Z$ and any sequence $\{\Delta^t\} \subset \mathcal{E}$, if we define $w_1^v = w_1$ and*

$$w_{i+1}^v = \underset{u\in Z}{argmin} -\gamma_i\langle \Delta^i, u\rangle + V(w_i^v, u), \quad \forall i > 1, \tag{54}$$

then

$$\sum_{i=1}^t \theta_i\langle -\Delta^i, w_i^v - u\rangle \le \sum_{i=1}^t \frac{\theta_i}{\gamma_i}(V(w_i^v, u) - V(w_{i+1}^v, u)) + \sum_{i=1}^t \frac{\theta_i\gamma_i}{2\mu}\|\Delta_i\|_*^2, \quad \forall u \in Z. \tag{55}$$

Proof Applying Lemma 1 to (54) (with $r = w_i^v$, $w = w_{i+1}^v$, $\zeta = -\gamma_i\Delta^i$ and $J = 0$), we have

$$-\gamma_i\langle \Delta^i, w_{i+1}^v - u\rangle \le V(w_i^v, u) - V(w_i^v, w_{i+1}^v) - V(w_{i+1}^v, u), \quad \forall u \in Z.$$

Moreover, by Schwartz inequality, Young's inequality and (46) we have

$$-\gamma_i\langle\Delta^i, w_i^v - w_{i+1}^v\rangle \le \gamma_i\|\Delta^i\|_*\|w_i^v - w_{i+1}^v\| \le \frac{\gamma_i^2}{2\mu}\|\Delta_i\|_*^2 + \frac{\mu}{2}\|w_i^v - w_{i+1}^v\|^2$$
$$\le \frac{\gamma_i^2}{2\mu}\|\Delta_i\|_*^2 + V(w_i^v, w_{i+1}^v).$$

Adding the above two inequalities and multiplying the resulting inequality by θ_i/γ_i, we obtain

$$-\theta_i\langle\Delta^i, w_i^v - u\rangle \le \frac{\theta_i\gamma_i}{2\mu}\|\Delta_i\|_*^2 + \frac{\theta_i}{\gamma_i}(V(w_i^v, u) - V(w_{i+1}^v, u)).$$

Summing the above inequalities from $i = 1$ to t, we conclude (55).

With the help of Lemma 4 and 5, we are now ready to prove Theorem 1, which provides an estimate of the gap function of SAMP in both expectation and probability.

Proof of Theorem 1 We first provide a bound on $\mathcal{B}_t(u, r_{[t]})$. Since the sequence $\{r_i\}_{i=1}^{t+1}$ is in the bounded set Z, applying (19) and (20) to (51) we have

$$\begin{aligned}
&\mathcal{B}_t(u, r_{[t]}) \\
&= \frac{\alpha_1}{\Gamma_1\gamma_1}V(r_1, u) - \sum_{i=1}^{t-1}\left[\frac{\alpha_i}{\Gamma_i\gamma_i} - \frac{\alpha_{i+1}}{\Gamma_{i+1}\gamma_{i+1}}\right]V(r_{t+1}[i], u) - \frac{\alpha_t}{\Gamma_t\gamma_t}V(r_{t+1}, u) \\
&\le \frac{\alpha_1}{\Gamma_1\gamma_1}\Omega_Z^2 - \sum_{i=1}^{t-1}\left[\frac{\alpha_i}{\Gamma_i\gamma_i} - \frac{\alpha_{i+1}}{\Gamma_{i+1}\gamma_{i+1}}\right]\Omega_Z^2 = \frac{\alpha_t}{\Gamma_t\gamma_t}\Omega_Z^2, \ \forall u \in Z,
\end{aligned} \tag{56}$$

Applying (20) and the above inequality to (50) in Lemma 4, we have

$$\frac{1}{\Gamma_t}Q(w_{t+1}^{ag}, u) \le \frac{\alpha_t}{\Gamma_t\gamma_t}\Omega_Z^2 + \sum_{i=1}^{t}\Lambda_i(u), \quad \forall u \in Z. \tag{57}$$

Letting $w_1^v = w_1$, defining w_{i+1}^v as in (54) with $\Delta^i = \Delta_H^{2i} + \Delta_G^i$ for all $i > 1$, we conclude from (51) and Lemma 5 (with $\theta_i = \alpha_i/\Gamma_i$) that

$$-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i} + \Delta_G^i, w_i^v - u\rangle \le \mathcal{B}_t(u, w_{[t]}^v) + \sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i} + \Delta_G^i\|_*^2, \quad \forall u \in Z. \tag{58}$$

The above inequality together with (52) and the Young's inequality yield

$$\begin{aligned}\sum_{i=1}^{t}\Lambda_i(u) &= -\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle \Delta_H^{2i}+\Delta_G^i, w_i^v-u\rangle + \sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2+\|\Delta_H^{2i-1}\|_*^2\right)\\ &\quad+\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\left[-\frac{(1-q)\mu}{2\gamma_i}\|r_i-w_{i+1}\|^2-\langle\Delta_G^i, w_{i+1}-r_i\rangle\right]\\ &\quad-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_G^i, r_i-w_i^v\rangle-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}, w_{i+1}-w_i^v\rangle\\ &\le \mathcal{B}_t(u, w_{[t]}^v)+U_t,\end{aligned} \tag{59}$$

where

$$\begin{aligned}U_t &:= \sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i}+\Delta_G^i\|_*^2+\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2(1-q)\mu\Gamma_i}\|\Delta_G^i\|_*^2\\ &\quad+\sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2+\|\Delta_H^{2i-1}\|_*^2\right)\\ &\quad-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_G^i, r_i-w_i^v\rangle-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}, w_{i+1}-w_i^v\rangle.\end{aligned} \tag{60}$$

Applying (56) and (59) to (57), we have

$$\frac{1}{\Gamma_t}Q(w_{t+1}^{ag}, u)\le\frac{2\alpha_t}{\gamma_t\Gamma_t}\Omega_Z^2+U_t,\quad\forall u\in Z,$$

or equivalently,

$$g(w_{t+1}^{ag})\le\frac{2\alpha_t}{\gamma_t}\Omega_Z^2+\Gamma_t U_t. \tag{61}$$

Now it suffices to bound U_t, in both expectation and probability.

We prove part (a) first. By our assumptions on $\mathcal{SO}_G$ and $\mathcal{SO}_H$ and in view of (13), (14) and (54), during the i-th iteration of Algorithm 1, the random noise Δ_H^{2i} is independent of w_{i+1} and w_i^v, and Δ_G^i is independent of r_i and w_i^v, hence $\mathbb{E}[\langle\Delta_G^i, r_i-w_i^v\rangle]=\mathbb{E}[\langle\Delta_H^{2i}, w_{i+1}-w_i^v\rangle]=0$. In addition, Assumption A1 implies that $\mathbb{E}[\|\Delta_G^i\|_*^2]\le\sigma_G^2$, $\mathbb{E}[\|\Delta_H^{2i-1}\|_*^2]\le\sigma_H^2$ and $\mathbb{E}[\|\Delta_H^{2i}\|_*^2]\le\sigma_H^2$, where Δ_G^i, Δ_H^{2i-1} and Δ_H^{2i} are independent. Therefore, taking expectation on (60) we have

$$\begin{aligned}\mathbb{E}[U_t] &\le \mathbb{E}\left[\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\left(\|\Delta_H^{2i}\|^2+\|\Delta_G^i\|_*^2\right)+\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2(1-q)\mu\Gamma_i}\|\Delta_G^i\|_*^2\right.\\ &\quad\left.+\sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2+\|\Delta_H^{2i-1}\|_*^2\right)\right] \qquad (62)\\ &=\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\left[4\sigma_H^2+\left(1+\frac{1}{2(1-q)}\right)\sigma_G^2\right].\end{aligned}$$

Taking expectation on both sides of (61), and using (62), we obtain (21).

Next we prove part (b). Observe that the sequence $\{\langle\Delta_G^i, r_i-w_i^v\rangle\}_{i\ge1}$ is a martingale difference and hence satisfies the large-deviation theorem (see, e.g., Lemma 2 of [25]). Therefore using Assumption A2 and the fact that

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{\frac{\mu(\alpha_i\Gamma_i^{-1}\langle\Delta_G^i,r_i-w_i^v\rangle)^2}{2(\sigma_G\alpha_i\Gamma_i^{-1}\Omega_Z)^2}\right\}\right]\\ &\le\ \mathbb{E}\left[\exp\left\{\frac{\mu\|\Delta_G^i\|_*^2\|r_i-w_i^v\|^2}{2\sigma_G^2\Omega_Z^2}\right\}\right]\le\mathbb{E}\left[\exp\left\{\|\Delta_G^i\|_*^2/\sigma_G^2\right\}\right]\le\exp\{1\},\end{aligned}$$

we conclude from the large-deviation theorem that

$$Prob\left\{-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_G^i,r_i-w_i^v\rangle>\lambda\sigma_G\Omega_Z\sqrt{\frac{2}{\mu}\sum_{i=1}^{t}\left(\frac{\alpha_i}{\Gamma_i}\right)^2}\right\}\le\exp\{-\lambda^2/3\}. \qquad (63)$$

By using a similar argument we have

$$Prob\left\{-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i},w_{i+1}-w_i^v\rangle>\lambda\sigma_H\Omega_Z\sqrt{\frac{2}{\mu}\sum_{i=1}^{t}\left(\frac{\alpha_i}{\Gamma_i}\right)^2}\right\}\le\exp\{-\lambda^2/3\}. \qquad (64)$$

In addition, letting $S_i=\alpha_i\gamma_i/(\mu\Gamma_i)$ and $S=\sum_{i=1}^t S_i$, by Assumption A2 and the convexity of exponential functions, we have

$$\mathbb{E}\left[\exp\left\{\frac{1}{S}\sum_{i=1}^{t}S_i\|\Delta_G^i\|_*^2/\sigma_G^2\right\}\right]\le\mathbb{E}\left[\frac{1}{S}\sum_{i=1}^{t}S_i\exp\left\{\|\Delta_G^i\|_*^2/\sigma_G^2\right\}\right]\le\exp\{1\}.$$

Noting by Markov's inequality that $P(X>a)\le\mathbb{E}[X]/a$ for all nonnegative random variables X and constants $a>0$, the above inequality implies that

$$Prob\left[\sum_{i=1}^{t}S_i\|\Delta_G^i\|_*^2>(1+\lambda)\sigma_G^2S\right]$$

$$
\begin{aligned}
&= Prob\left[\exp\left\{\frac{1}{S}\sum_{i=1}^{t} S_i \|\Delta_G^i\|_*^2/\sigma_G^2\right\} > \exp\{1+\lambda\}\right] \\
&\le \mathbb{E}\left[\exp\left\{\frac{1}{S}\sum_{i=1}^{t} S_i \|\Delta_G^i\|_*^2/\sigma_G^2\right\}\right] \Big/ \exp\{1+\lambda\} \\
&\le \exp\{-\lambda\}.
\end{aligned}
$$

Recalling that $S_i = \alpha_i\gamma_i/(\mu\Gamma_i)$ and $S = \sum_{i=1}^t S_i$, the above relation is equivalent to

$$
\begin{aligned}
&Prob\left\{\left(1+\frac{1}{2(1-q)}\right)\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\|\Delta_G^i\|_*^2 > (1+\lambda)\sigma_G^2\left(1+\frac{1}{2(1-q)}\right)\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\right\} \\
&\le \exp\{-\lambda\}.
\end{aligned} \tag{65}
$$

Using similar arguments, we also have

$$
Prob\left\{\sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i-1}\|_*^2 > (1+\lambda)\frac{3\sigma_H^2}{2}\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\right\} \le \exp\{-\lambda\}, \tag{66}
$$

$$
Prob\left\{\sum_{i=1}^{t}\frac{5\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i}\|_*^2 > (1+\lambda)\frac{5\sigma_H^2}{2}\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\right\} \le \exp\{-\lambda\}. \tag{67}
$$

Using the fact that $\|\Delta_H^{2i} + \Delta_G^{2i-1}\|_*^2 \le 2\|\Delta_H^{2i}\|_*^2 + 2\|\Delta_G^{2i-1}\|_*^2$, we conclude from (61)–(67) that (22) holds.

In the remaining part of this subsection, we will focus on proving Theorem 2, which describes the rate of convergence of Algorithm 1 for solving $SVI(Z; G, H, J)$ when Z is unbounded.

Proof the Theorem 2 Let U_t be defined in (60). Firstly, applying (29) and (59) to (50) in Lemma 4, and noting that $\mu = 1$, we have

$$
\frac{1}{\Gamma_t}Q(w_{t+1}^{ag}, u) \tag{68}
$$

$$
\le \mathcal{B}_t(u, r_{[t]}) - \frac{\alpha_t}{2\Gamma_t\gamma_t}\sum_{i=1}^{t}\left(q-c^2\right)\|r_i - w_{i+1}\|^2 + \mathcal{B}_t(u, w_{[t]}^v) + U_t, \quad \forall u \in Z. \tag{69}
$$

In addition, applying (29) to the definition of $\mathcal{B}_t(\cdot,\cdot)$ in (51), we obtain

$$\mathcal{B}_t(u, r_{[t]}) = \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1 - u\|^2 - \|r_{t+1} - u\|^2) \tag{70}$$

$$= \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1 - w_{t+1}^{ag}\|^2 - \|r_{t+1} - w_{t+1}^{ag}\|^2 + 2\langle r_1 - r_{t+1}, w_{t+1}^{ag} - u\rangle). \tag{71}$$

By using a similar argument and the fact that $w_1^v = w_1 = r_1$, we have

$$\mathcal{B}_t(u, w_{[t]}^v) = \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1 - u\|^2 - \|w_{t+1}^v - u\|^2) \tag{72}$$

$$= \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1 - w_{t+1}^{ag}\|^2 - \|w_{t+1}^v - w_{t+1}^{ag}\|^2 + 2\langle r_1 - w_{t+1}^v, w_{t+1}^{ag} - u\rangle). \tag{73}$$

We then conclude from (68), (71), and (73) that

$$Q(w_{t+1}^{ag}, u) - \langle v_{t+1}, w_{t+1}^{ag} - u\rangle \le \varepsilon_{t+1}, \quad \forall u \in Z, \tag{74}$$

where

$$v_{t+1} := \frac{\alpha_t}{\gamma_t}(2r_1 - r_{t+1} - w_{t+1}^v) \tag{75}$$

and

$$\varepsilon_{t+1} := \frac{\alpha_t}{2\gamma_t}\left(2\|r_1 - w_{t+1}^{ag}\|^2 - \|r_{t+1} - w_{t+1}^{ag}\|^2 - \|w_{t+1}^v - w_{t+1}^{ag}\|^2 - \sum_{i=1}^{t}\left(q - c^2\right)\|r_i - w_{i+1}\|^2\right) + \Gamma_t U_t. \tag{76}$$

It is easy to see that the residual ε_{t+1} is positive by setting $u = w_{t+1}^{ag}$ in (74). Hence $\tilde{g}(w_{t+1}^{ag}, v_{t+1}) \le \varepsilon_{t+1}$. To finish the proof, it suffices to estimate the bounds for $\mathbb{E}[\|v_{t+1}\|]$ and $\mathbb{E}[\varepsilon_{t+1}]$. Observe that by (2), (6), (16) and the convexity of G and J, we have

$$Q(w_{t+1}^{ag}, u^*) \ge \langle F(u^*), w_{t+1}^{ag} - u^*\rangle \ge 0, \tag{77}$$

where the last inequality follows from the assumption that u^* is a strong solution of $SVI(Z; G, H, J)$. Using the above inequality and letting $u = u^*$ in (68), we conclude from (70) and (72) that

$$2\|r_1 - u^*\|^2 - \|r_{t+1} - u^*\|^2 - \|w_{t+1}^v - u^*\|^2 - \sum_{i=1}^{t}\left(q - c^2\right)\|r_i - w_{i+1}\|^2 + \frac{2\Gamma_t\gamma_t}{\alpha_t}U_t \geq \frac{2\gamma_t}{\alpha_t}Q(w_{t+1}^{ag}, u^*) \geq 0.$$

By the above inequality and the definition of D in (32), we have

$$\|r_{t+1} - u^*\|^2 + \|w_{t+1}^v - u^*\|^2 + \sum_{i=1}^{t}\left(q - c^2\right)\|r_i - w_{i+1}\|^2 \leq 2D^2 + \frac{2\Gamma_t\gamma_t}{\alpha_t}U_t. \tag{78}$$

In addition, applying (29) and the definition of C_t in (33) to (62), we have

$$\mathbb{E}[U_t] \leq \sum_{i=1}^{t}\frac{\alpha_t\gamma_i^2}{\Gamma_t\gamma_t}\left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right] = \frac{\alpha_t}{\Gamma_t\gamma_t}C_t^2. \tag{79}$$

Combining (78) and (79), we have

$$\mathbb{E}[\|r_{t+1} - u^*\|^2] + \mathbb{E}[\|w_{t+1}^v - u^*\|^2] + \sum_{i=1}^{t}\left(q - c^2\right)\mathbb{E}[\|r_i - w_{i+1}\|^2] \leq 2D^2 + 2C_t^2. \tag{80}$$

We are now ready to prove (30). Observe from the definition of v_{t+1} in (75) and the definition of D in (32) that $\|v_{t+1}\| \leq \alpha_t(2D + \|w_{t+1}^v - u^*\| + \|r_{t+1} - u^*\|)/\gamma_t$, using the previous inequality, Jensen's inequality, and (80), we obtain

$$\begin{aligned}\mathbb{E}[\|v_{t+1}\|] &\leq \frac{\alpha_t}{\gamma_t}\left(2D + \sqrt{\mathbb{E}[(\|r_{t+1} - u^*\| + \|w_{t+1}^v - u^*\|)^2]}\right)\\ &\leq \frac{\alpha_t}{\gamma_t}\left(2D + \sqrt{2\mathbb{E}[\|r_{t+1} - u^*\|^2 + \|w_{t+1}^v - u^*\|^2]}\right)\\ &\leq \frac{\alpha_t}{\gamma_t}\left(2D + 2\sqrt{D^2 + C_t^2}\right).\end{aligned}$$

Our remaining goal is to prove (31). By (15) and (18), we have

$$\frac{1}{\Gamma_t}w_{t+1}^{ag} = \frac{1}{\Gamma_{t-1}}w_t^{ag} + \frac{\alpha_t}{\Gamma_t}w_{t+1}, \quad \forall t > 1.$$

Using the assumption that $w_1^{ag} = w_1$, we obtain

$$w_{t+1}^{ag} = \Gamma_t\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}w_{i+1}, \tag{81}$$

where by (18) we have

$$\Gamma_t \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i} = 1. \tag{82}$$

Therefore, w_{t+1}^{ag} is a convex combination of iterates $w_2, \ldots, w_{t+1}$. Also, by a similar argument in the proof of Lemma 4, applying Lemma 2 to (13) and (14) (with $r = r_t$, $w = w_{t+1}$, $y = r_{t+1}$, $\eta = \gamma_t \mathcal{H}(r_t; \zeta_{2t-1}) + \gamma_t \mathcal{G}(w_t^{md}; \xi_t)$, $\vartheta = \gamma_t \mathcal{H}(w_{t+1}; \zeta_{2t}) + \gamma_t \mathcal{G}(w_t^{md}; \xi_t)$, $J = \gamma_t J$, $L = 3M^2\gamma_t^2$ and $M^2 = 3\gamma_t^2(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2)$), and using (42) and (53), we have

$$\begin{aligned}\frac{1}{2}\|r_{t+1} - w_{t+1}\|^2 &\le \frac{3M^2\gamma_t^2}{2}\|r_t - w_{t+1}\|^2 + \frac{3\gamma_t^2}{2}(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2) \\ &\le \frac{c^2}{2}\|r_t - w_{t+1}\|^2 + \frac{3\gamma_t^2}{2}(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2),\end{aligned}$$

where the last inequality follows from (29).

Now using (76), (81), (82), the above inequality, and applying Jensen's inequality, we have

$$\begin{aligned}\varepsilon_{t+1} - \Gamma_t U_t \le \frac{\alpha_t}{\gamma_t}\|r_1 - w_{t+1}^{ag}\|^2 &= \frac{\alpha_t}{\gamma_t}\left\|r_1 - u^* + \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}(u^* - r_{i+1})\right. \\ &\quad \left. + \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}(r_{i+1} - w_{i+1})\right\|^2 \\ &\le \frac{3\alpha_t}{\gamma_t}\left[D^2 + \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\left(\|r_{i+1} - u^*\|^2 + \|w_{i+1} - r_{i+1}\|^2\right)\right] \\ &\le \frac{3\alpha_t}{\gamma_t}\left[D^2 + \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\Big(\|r_{i+1} - u^*\|^2 + c^2\|w_{i+1} - r_i\|^2\right. \\ &\quad \left. + 3\gamma_i^2(\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2)\|\Big)\right].\end{aligned} \tag{83}$$

Noting that by (33) and (78),

$$\begin{aligned}&\Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}(\|r_{i+1} - u^*\|^2 + c^2\|w_{i+1} - r_i\|^2) \\ &\le \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i\theta}{\Gamma_i}(\|r_{i+1} - u^*\|^2 + (q - c^2)\|w_{i+1} - r_i\|^2) \\ &\le \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i\theta}{\Gamma_i}(2D^2 + \frac{2\Gamma_i\gamma_i}{\alpha_i}U_i) = 2\theta D^2 + 2\theta\Gamma_t \sum_{i=1}^{t}\gamma_i U_i,\end{aligned}$$

and that by (29),

$$\Gamma_t \sum_{i=1}^{t} \frac{3\alpha_i \gamma_i^2}{\Gamma_i}(\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2) = \Gamma_t \sum_{i=1}^{t} \frac{3\alpha_t \gamma_i^3}{\Gamma_t \gamma_t}(\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2)$$
$$= \frac{3\alpha_t}{\gamma_t} \sum_{i=1}^{t} \gamma_i^3 (\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2),$$

we conclude from (79), (83) and Assumption A1 that

$$\mathbb{E}[\varepsilon_{t+1}] \le \Gamma_t \mathbb{E}[U_t] + \frac{3\alpha_t}{\gamma_t}\left[D^2 + 2\theta D^2 + 2\theta \Gamma_t \sum_{i=1}^{t} \gamma_i \mathbb{E}[U_i] + \frac{6\alpha_t \sigma_H^2}{\gamma_t} \sum_{i=1}^{t} \gamma_i^3\right]$$
$$\le \frac{\alpha_t}{\gamma_t} C_t^2 + \frac{3\alpha_t}{\gamma_t}\left[(1+2\theta)D^2 + 2\theta \Gamma_t \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i} C_i^2 + \frac{6\alpha_t \sigma_H^2}{\gamma_t} \sum_{i=1}^{t} \gamma_i^3\right].$$

Finally, observing from (33) and (82) that

$$\Gamma_t \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i} C_i^2 \le C_t^2 \Gamma_t \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i} = C_t^2,$$

we conclude (31) from the above inequality.

4 Numerical experiments

In this section, we present some preliminary experimental results on solving deterministic and stochastic variational inequality problems using the SAMP algorithm. The comparisons with the mirror-prox method in [32] and the stochastic mirror-prox method in [20] are provided for better examination of the performance of the SAMP algorithm.

4.1 Overlapped group lasso

Our first numerical experiment is on a problem of form (7). Specifically, we consider the following overlapped group lasso problem:

$$\min_{x \in X} \frac{1}{2} \mathbb{E}_{a,f}[(\langle a, x\rangle - f)^2] + \lambda \sum_{g \in \mathcal{S}} \|x_g\|. \tag{84}$$

Here, the feasible set X is a Euclidean ball with $X := \{x \in \mathbb{R}^n | \|x\| \le D\}$, $\|\cdot\|$ is the Euclidean norm, the random variable pair (a, f) represents a dataset of interest, and x is the sparse feature of the dataset to be extracted. The sparsity structure of x is represented by group $\mathcal{S} \subseteq 2^{\{1,\ldots,n\}}$, and for any $g \subseteq \{1, \ldots, n\}$, x_g is a sparse vector

that is constructed by components of x whose indices are in g, i.e., $x_g := (x_i)_{i\in g}$ and there are very few number of non-zero components in x_g. Problems utilizing such sparsity structure is known as the overlapped group lasso [18]. In this experiment, we assume that each group $g \in \mathcal{S}$ consists of k elements. The first term in (84) describes the fidelity of the relation between dataset and its underlying feature, and the second term is the regularization term to enforce certain group sparsity. Problem (84) can be formulated as a SVI problem (1) with $\xi = (a, f)$ and

$$\mathcal{F}(u;\xi) = \begin{pmatrix} (\langle a, x\rangle - f)a + K^T y \\ -K^T x \end{pmatrix}, \quad \forall u = (x, y) \in Z := X \times Y. \tag{85}$$

Here the linear operator K is defined by $Kx = \lambda(x_{g_1}^T, x_{g_2}^T, \ldots, x_{g_l}^T)^T$, where $g_i \in \mathcal{S}$ and $\mathcal{S} = \{g_i\}_{i=1}^l$. The set X is $\mathbb{R}^n$, and the set Y is the set of vectors $y \in \mathbb{R}^{kn}$ that are composed by n sub-vectors in the k dimensional unit ball:

$$Y := \left\{ y \in \mathbb{R}^{kn} \,|\, \left\| \left(y^{(ki-k+1)}, y^{(ki-k+2)}, \ldots, y^{(ki)} \right)^T \right\| \le 1 \right\}. \tag{86}$$

We can see that F has the form of (2) where

$$\mathcal{G}(u;\xi) = \begin{pmatrix} \langle (a, x\rangle - f)a \\ 0 \end{pmatrix}, \ \mathcal{H}(u;\xi) = H(u) = \begin{pmatrix} 0 & K^T \\ -K & 0 \end{pmatrix} u, \text{ and } J(u) \equiv 0. \tag{87}$$

In this experiment, we assume that the k-th components of a satisfy $a^{(k)} \sim N(0, 1)$ for all k, $f = \langle a, x_{true}\rangle + \varepsilon$ for some underlying ground truth x_{true} and noise $\varepsilon \sim N(0, \sigma_{noise}^2)$. The goal of solving (84) is to recover a feature x that is as close to x_{true} as possible, with only the knowledge of norm $D := \|x_{true}\|$ of the underlying ground truth. With such assumptions, it can be computed that $G(u) = \|x - x_{true}\|^2/2 + \sigma_{noise}^2$. The stochastic gradients $\mathcal{G}(u;\xi)$ is computed through a mini-batch fashion with batch size b, namely, $\mathcal{G}(u;\xi) = (\sum_{i=1}^b (\langle a_i, x\rangle - f_i)a_i)/b$ with b independently generated samples (a_i, f_i), where $a_i^{(k)} \sim N(0, 1)$ for all the k-th components of $a^{(k)}$, $f_i = \langle a_i, x_{true}\rangle + \varepsilon_i$, and $\varepsilon_i \sim N(0, \sigma_{noise}^2)$. Fixing any $x \in X$ and denoting that $d := x - x_{true}$ and $\kappa_i := (\langle a_i, d\rangle - \varepsilon_i)a_i - d$, we have $\mathbb{E}[\kappa_i] = 0$ and $\|d\| \le 2D$, hence

$$\begin{aligned}
\mathbb{E}\left[\|\kappa_i\|^2\right] &= \mathbb{E}\left[\|(\langle a_i, d\rangle - \varepsilon_i)a_i\|^2\right] - \|d\|^2 \\
&= \mathbb{E}\left[\|\langle a_i, d\rangle a_i\|\right]^2 + \mathbb{E}\left[\varepsilon_i^2\|a_i\|^2\right] - \|d\|^2 \\
&= \mathbb{E}\left[\sum_{j=1}^n \left(a_i^{(j)} d^{(j)} + \sum_{\substack{k=1 \\ k\neq j}}^n a_i^{(k)} d^{(k)} \right)^2 \left(a_i^{(j)}\right)^2\right] + n\sigma_{noise}^2 - \|d\|^2 \\
&= \mathbb{E}\left[\sum_{j=1}^n \left(\left(a_i^{(j)}\right)^2 d^{(j)} + \sum_{\substack{k=1 \\ k\neq j}}^n a_i^{(j)} a_i^{(k)} d^{(k)} \right)^2\right] + n\sigma_{noise}^2 - \|d\|^2
\end{aligned}$$

$$= \mathbb{E}\left[\sum_{j=1}^{n}\left(\left(a_i^{(j)}\right)^4\left(d^{(j)}\right)^2 + \sum_{\substack{k=1\\k\neq j}}^{n}\left(a_i^{(j)}a_i^{(k)}\right)^2\left(d^{(k)}\right)^2\right)\right] + n\sigma_{noise}^2 - \|d\|^2$$

$$= \sum_{j=1}^{n}\left(3\left(d^{(j)}\right)^2 + \sum_{\substack{k=1\\k\neq j}}^{n}\left(d^{(k)}\right)^2\right) + n\sigma_{noise}^2 - \|d\|^2$$

$$= (n+1)\|d\|^2 + n\sigma_{noise}^2 \leq 4(n+1)D^2 + n\sigma_{noise}^2.$$

Therefore, noting that κ_i are i.i.d., we have

$$\mathbb{E}[\mathcal{G}(u;\xi)] = \mathbb{E}\left[\frac{1}{b}\sum_{i=1}^{b}\kappa_i + d\right] = x - x_{true} = \nabla G(u),$$

and

$$\mathbb{E}\left[\|\mathcal{G}(u;\xi) - \nabla G(u)\|_*^2\right] = \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\kappa_i\right\|^2\right] = \frac{1}{b^2}\sum_{i=1}^{b}\mathbb{E}[\|\kappa_i\|^2]$$
$$\leq \frac{4(n+1)D^2 + \sigma_{noise}^2 n}{b}.$$

By the above analysis, it can be computed that $L = 1$ in (4), $M = \|K\|$ in (3), $\sigma_G^2 = (4(n+1)D^2 + \sigma_{noise}^2 n)/b$ and $\sigma_H^2 = 0$ in Assumption A1. The true feature x_{true} is the n-vector form of a 64×64 two-dimensional signal whose intensities are shown in Fig. 1. Within its support, the nonzero intensities of x_{true} are generated independently from standard normal distribution. The group sparsity structure we enforce is a grid structure as described in [18] with all the 4-cycles, in order to enforce that each pixel in the support x_{true} is connected to the pixels that are above, below, left, and right of itself.

Fig. 1 True feature x_{true} in the experiment of overlapped group LASSO. From *left to right*: problem instances 1 through 3

Table 2 The comparison of the SAMP and SMP algorithms in extracting overlapped group sparse feature of datasets, in terms of the relative error (88) to the ground truth

Problem instance	Iteration	SAMP		SMP	
	N	Rel. error (%)	CPU	Rel. error (%)	CPU
1	500	26.8	1.7	89.3	3.4
	1000	11.6	3.5	80.1	6.8
	1500	6.5	5.3	72.2	10.2
	2000	4.7	7.0	65.3	13.7
2	500	26.8	1.9	89.3	3.3
	1000	11.6	3.8	80.1	6.6
	1500	6.5	5.7	72.2	9.9
	2000	4.7	7.7	65.3	13.3
3	500	26.7	1.7	89.3	3.3
	1000	11.6	3.5	80.1	6.6
	1500	6.5	5.3	72.1	9.9
	2000	4.7	7.1	65.3	13.2

The relative error and CPU time in the table is the average results of 100 runs

We present in Table 2 the comparison results between the SAMP algorithm and the stochastic mirror-prox (SMP) algorithm in [20]. Three problem instances are generated, and in all the instances we set $\lambda = 10^{-4}$ in (84), and $\sigma_{noise} = 0.1$ and $b = 50$ in the sampling process. The accuracy of feature extraction of algorithm output x is evaluated by the relative error to the ground truth, which is defined by

$$\frac{\|x - x_{true}\|}{\|x_{true}\|}. \tag{88}$$

For both the SAMP and SMP algorithms, the parameters are selected based on the recommended optimal settings. In particular, for the SAMP algorithm, we use the parameters described in Corollary 1 (in which $\beta = \sigma/D$). For the SMP algorithm, we use the parameters described in Corollary 1 (with $L = 1 + \|K\|$, $M = 0$, $\Omega = D$, and $\sigma = \sigma_G$) in [20] . We can observe that the SAMP algorithm significantly outperforms the SMP algorithm in extracting the underlying feature of the datasets.

4.2 Randomized algorithm for solving two-player game

The goal of this subsection is to demonstrate the efficiency of the SAMP algorithm in computing the equilibrium of a two-player game. In particular, we consider the SVI problem

$$\left\langle \mathbb{E}\left[\begin{pmatrix} \mathcal{P}(x;\xi_x) + \mathcal{K}_y(y;\zeta_y) \\ -\mathcal{K}_x(x;\zeta_x) + \mathcal{Q}(y;\xi_y) \end{pmatrix}\right], \begin{pmatrix} x^* - x \\ y^* - y \end{pmatrix} \right\rangle \leq 0, \quad \forall x, y \in \Delta^n, \tag{89}$$

where $\xi_x, \xi_y, \zeta_x, \zeta_y$ are random variables such that

$$\text{Prob}(\mathcal{P}(x;\xi_x) = P_j) = x_i^{(j)},\ \text{Prob}(\mathcal{Q}(y;\xi_y) = Q_k) = y_i^{(k)}, \tag{90}$$

$$\text{Prob}(\mathcal{K}_x(x;\zeta_x) = K_l) = x_i^{(l)},\ \text{and}\ \text{Prob}(\mathcal{K}_y(y;\zeta_y) = K^m) = y_i^{(m)}. \tag{91}$$

Here Δ^n is a standard simplex, P_j and Q_k are the j-th and k-th columns of positive semidefinite matrices P and Q, and K_l and $(K^m)^T$ are the l-th column and m-th row of a matrix K, respectively. Letting $Z := \Delta^n \times \Delta^n$, $\xi := (\xi_x, \xi_y)$, $\zeta := (\zeta_x, \zeta_y)$, and using the notation $u = (x, y) \in Z$, problem (89) can be formulated as (1) where

$$\mathcal{G}(u;\xi) = \begin{pmatrix} \mathcal{P}(x;\xi_x) \\ \mathcal{Q}(y;\xi_y) \end{pmatrix},\ \mathcal{H}(u;\zeta) = \begin{pmatrix} \mathcal{K}_y(y;\zeta_y) \\ -\mathcal{K}_x(x;\zeta_x) \end{pmatrix},\ \text{and}\ J(u) \equiv 0. \tag{92}$$

Also, it can be checked from (90) that $\mathbb{E}[\mathcal{P}(x;\xi_x)] = Px$, $\mathbb{E}[\mathcal{Q}(y;\xi_y)] = Qy$, $\mathbb{E}[\mathcal{K}_x(x;\zeta_x)] = Kx$ and $\mathbb{E}[\mathcal{K}_y(y;\zeta_y)] = K^T y$, hence problem (89) also have the equivalent form (5) where

$$F(u) = \begin{pmatrix} P & K^T \\ -K & Q \end{pmatrix} u,\ G(u) = \frac{1}{2}\langle Px, x\rangle + \frac{1}{2}\langle Qy, y\rangle,\ \text{and}\ H(u) = \begin{pmatrix} K^T y \\ -Kx \end{pmatrix},$$

or a saddle point problem

$$\min_{x\in\Delta^n} \max_{y\in\Delta^n} \frac{1}{2}\langle Px, x\rangle + \langle Kx, y\rangle - \frac{1}{2}\langle Qy, y\rangle. \tag{93}$$

The above saddle point problem describes the equilibrium of a two-player game.

It should be noted that, although problems (89) and (93) are equivalent, from the algorithm design point of view the stochastic formulation (89) may have some advantages over its deterministic counterpart (93), as pointed out in [33]. This is because that when P, Q and K are dense and n is large, the matrix-vector multiplication of Px, Qy, $K^T y$ and Kx may be relatively expensive. Consequently, the stochastic formulation (89) becomes more favorable than its deterministic counterpart (89), since each sampling of $\mathcal{P}(x;\xi_x)$, $\mathcal{Q}(y;\xi_y)$, $\mathcal{K}_y(y;\zeta_y)$ and $\mathcal{K}_x(x;\zeta_x)$ is just a random selection of a column or row, rather than a matrix-vector computation. Indeed, designing a stochastic approximation algorithm for solving the SVI (89) is equivalent to designing a randomized algorithm for solving (93).

To compute a solution of (1), we consider an entropy setting for the prox-function used in the AMP algorithm. For simplicity, we only consider in this experiment the case when $\max_{i,j} |P^{(i,j)}| = \max_{i,j} |Q^{(i,j)}|$.[1] For all $z = (\tilde{x}, \tilde{y}) \in Z$, $u = (x, y) \in Z$ and $\eta = (\eta_x, \eta_y) \in \mathcal{E}$, we define

[1] When the maximum absolute values of P and Q are different, it is recommended to introduce weights ω_x and ω_y and set $\|u\| := \sqrt{\omega_x \|x\|_1^2 + \omega_y \|y\|_1^2}$ and $\|\eta\|_* := \sqrt{\|\eta_x\|_1^2/\omega_x + \|\eta_y\|_1^2/\omega_y}$. See "mixed setups" in Section 5 of [32] for the detailed derivations for best values of weights ω_x and ω_y.

$$\|u\| := \sqrt{\|x\|_1^2 + \|y\|_1^2 s},\ \|\eta\|_* := \sqrt{\|\eta_x\|_\infty^2 + \|\eta_y\|_\infty^2},\ \text{and}$$
$$V(z,u) := \sum_{j=1}^{n}(x_i^{(j)} + \nu/n)\ln\frac{x_i^{(j)} + \nu/n}{\tilde{x}_i^{(j)} + \nu/n} + \sum_{j=1}^{n}(y_i^{(j)} + \nu/n)\ln\frac{y_i^{(j)} + \nu/n}{\tilde{y}_i^{(j)} + \nu/n}. \quad (94)$$

Here, $y_i^{(j)}$ denotes the j-th entry of the strategy y_i, and ν is arbitrarily small (e.g., $\nu = 10^{-16}$). With the above setting, the optimization problem in the prox-mapping (10) can be efficiently solved within machine accuracy, and the strong convexity parameter of the prox-function $V(z,u)$ is $\mu = 1 + \nu$ (See [3] for details on the entropy prox-functions). It is easy to check that

$$L \le \max\left\{\max_{k,j}|P^{(k,j)}|, \max_{k,j}|Q^{(k,j)}|\right\},\ M \le \max_{k,j}|K^{(k,j)}|,\ \Omega_Z^2 = 2\left(1 + \frac{\nu}{n}\right)\ln\left(\frac{n}{\nu} + 1\right),$$
$$\mathbb{E}_\xi\left[\|\mathcal{G}(u;\xi) - \nabla G(u)\|_*^2\right] \le 4\left(\max_{k,j}|P^{(k,j)}|^2 + \max_{k,j}|Q^{(k,j)}|^2\right),\ \text{and}$$
$$\mathbb{E}_\zeta\left[\|\mathcal{H}(u;\zeta) - H(u)\|_*^2\right] \le 8\max_{k,j}|K^{(k,j)}|^2.$$

Therefore, we set

$$\sigma_G = 2\sqrt{\left(\max_{k,j}|P^{(k,j)}|^2 + \max_{k,j}|Q^{(k,j)}|^2\right)},\ \sigma_H = 2\sqrt{2\max_{k,j}|K^{(k,j)}|},$$

and σ by (8).

In this experiment, we generate random matrices $B, C \in \mathbb{R}^{100\times n}$, and $K \in \mathbb{R}^{n\times n}$ first, where each entry of these matrices are independently and uniformly distributed over [0, 1]. The matrices P and Q are then generated by $P = B^T B$ and $Q = C^T C$, and also rescaled so that P and Q are both positive semidefinite and $L = \max_{k,j}|P^{(k,j)}| = \max_{k,j}|Q^{(k,j)}|$. For the SAMP algorithm, we use the scheme in Algorithm 1 with the parameters described in (94) and Corollary 1 (in which $\beta = \sigma/\Omega_Z$). As a comparison, we also implement the stochastic mirror-prox (SMP) method in with parameters set by Corollay 1 in [20] (in which $L = \sqrt{2\max_{k,j}|P^{(k,j)}|^2 + 2\max_{k,j}|K^{(k,j)}|^2}$, $\sigma = 4\sqrt{\max_{k,j}|P^{(k,j)}|^2 + \max_{k,j}|K^{(k,j)}|^2}$ and $\Omega = \Omega_Z^2$). The performance of the SAMP and SMP algorithms are compared in terms of the average of the gap function values (17) (computed by MOSEK [30]) in 100 runs.

The comparison between the SAMP and SMP algorithms in terms of the performance on computing approximate solutions of (93) is described in Table 3. We can see that the SAMP algorithm outperforms the SMP algorithm, which is consistent with our theoretical observation on the iteration complexities of the SAMP and SMP algorithms. In particular, as L increases, the advantage of SAMP over SMP in terms of $\mathbb{E}[g(u)]$ becomes more evident.

Table 3 The comparison of the SAMP and SMP algorithms in solving SVI (89), in terms of the expectation of the gap function value $g(u)$ for any approximate solution u

Problem instance	Iteration N	SAMP $\mathbb{E}[g(u)]$	SAMP CPU (ave.)	SMP $\mathbb{E}[g(u)]$	SMP CPU (ave.)
1	1000	6.39e−2	0.4	7.69e−2	0.4
	2000	5.82e−2	0.7	7.27e−2	0.8
	5000	4.54e−2	1.7	6.43e−2	2.0
2	1000	4.09e−2	0.5	5.05e−1	0.6
	2000	1.86e−2	0.9	4.84e−1	1.1
	5000	6.78e−3	2.3	4.44e−1	2.8
3	1000	9.51e−2	0.9	3.72e0	1.1
	2000	5.81e−2	2.2	3.63e0	2.6
	5000	3.27e−2	5.4	3.45e0	6.6

The CPU time in the table is the average time of 100 runs
Instance 1: $n = 1000$, $L = 1$, $M = 1$, $\sigma = 4.00$
Instance 2: $n = 2000$, $L = 10$, $M = 1$, $\sigma = 9.38$
Instance 3: $n = 5000$, $L = 100$, $M = 1$, $\sigma = 28.43$

4.3 Two-player game with nonlinear payoff

Our goal in this subsection is to demonstrate the advantages of the SAMP algorithm over the mirror-prox method (or extragradient method) even for solving certain classes of deterministic VIs. We consider a problem on computing the equilibrium of a convex-concave two-player game of form

$$\min_{x\in\Delta^n}\max_{y\in\Delta^n}\sum_{i=1}^{m}\log\left(1+e^{\langle a_i,x\rangle}\right)+\sum_{i=1}^{n}\log\left(1+\frac{y^{(i)}}{c^{(i)}+x^{(i)}}\right)-\sum_{i=1}^{m}\log\left(1+e^{\langle b_i,x\rangle}\right), \tag{95}$$

where Δ^n is the standard simplex. When $a_i = b_i = 0$ for all $i = 1, \ldots, m$, the above becomes the water filling problem (see, e.g., [5]). Letting $Z := \Delta^n \times \Delta^n$ and using the notation $u = (x, y)$, the above problem is equivalent to (5) with $J(u) \equiv 0$ and

$$G(u) = \sum_{i=1}^{m}\log\left(1+e^{\langle a_i,x\rangle}\right)+\sum_{i=1}^{m}\log\left(1+e^{\langle b_i,x\rangle}\right), \tag{96}$$

$$H(u) = \begin{pmatrix} -\dfrac{y^{(i)}}{(c^{(i)}+x^{(i)})(c^{(i)}+x^{(i)}+y^{(i)})} \\ -\dfrac{1}{c^{(i)}+x^{(i)}+y^{(i)}} \end{pmatrix}. \tag{97}$$

In this experiment, we generate a_i and b_i randomly from the standard normal distribution, and set $c^{(i)} = 1$ for all i. We apply the entropy setting in (94). For

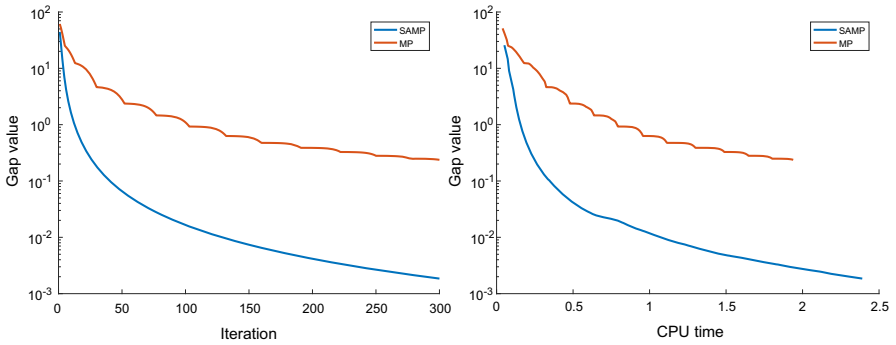

Fig. 2 Performance of SAMP and MP on solving problem (95). *Left*: iteration versus gap value $g(u)$. *Right*: CPU time versus gap value $g(u)$

the SAMP algorithm, we incorporate a backtracking linesearch technique in order to determine the best values of L and M that guarantees the convergence of Algorithm 1 (see, e.g., the backtracking technique in [36]). We compare the SAMP algorithm with the mirror-prox (MP) algorithm with adaptive stepsize described in [32]. The comparison between the SAMP algorithm and the MP algorithm is illustrated through the convergence of gap function $g(u)$ in (17) (computed by IPOPT [47]). It can be observed from Fig. 2 that the SAMP algorithm outperforms the MP algorithm in terms of both iteration vs. gap value and CPU time versus gap value. This is consistent with our theoretical observation that SAMP has a better iteration complexity bound than the MP algorithm for solving deterministic VIs.

4.4 Variational inequality on the Lorentz cone

In this section, we study the performance of SAMP when the feasible set is unbounded. In particular, we consider an SVI problem on solving $u^* \in Z$ such that

$$\langle \mathbb{E}[Au + \zeta], u^* - u \rangle \leq 0, \quad \forall u \in Z, \tag{98}$$

where $A \in \mathbb{R}^{(n+1)\times(n+1)}$ is a linear monotone operator, ξ is a random vector whose expectation $b = \mathbb{E}[\zeta]$ is unknown *a priori*, and Z is the Lorentz cone:

$$Z := \{(x, t) \in \mathbb{R}^{(n+1)} \mid \|x\| \leq t\}.$$

To solve (98), we can decompose the linear monotone operator A to the sum of a symmetric positive semidefinite matrix $(A + A^T)/2$ and a skew-symmetric matrix $(A - A^T)/2$, hence the SVI problem (98) can be viewed as an instance of (1) and (5) with

Table 4 The comparison of the SAMP and MP algorithms in solving the SVI problem (98)

Problem	N	SAMP, N iterations			MP, $2N$ iterations		
		$\mathbb{E}[\tilde{g}(w,v)]$	$\mathbb{E}[\|v\|]$	CPU	$\tilde{g}(w,v)$	$\|v\|$	CPU
$n = 1999$	1000	8.0e−2	3.6e−1	5.0	7.6e−2	3.3e0	6.7
$L = 5861.5$	2000	3.6e−2	1.6e−1	10.2	1.1e−1	2.3e0	13.9
$M = 36.5$	4000	2.3e−2	1.1e−1	15.3	1.2e−1	1.8e0	20.5
$n = 2999$	1000	4.5e−2	3.1e−1	10.7	6.0e−1	3.6e0	14.9
$L = 8645.0$	2000	2.0e−2	1.4e−1	22.2	7.6e−2	2.4e0	30.6
$M = 44.6$	3000	1.3e−2	9.3e−2	33.6	7.9e−2	1.8e0	45.6

In the table w and v denote the approximate solution and perturbation vector respectively, and $\tilde{g}(\cdot,\cdot)$ is the gap function defined in (28)
The CPU time in the table is the average time of 100 runs

$$F(u) = Au + b,\ G(u) = \frac{1}{4}\langle (A + A^T)u, u\rangle,\ H(u) = \frac{1}{2}(A - A^T)u + b,\ J(u) = 0,$$
$$\mathcal{G}(u;\xi) \equiv \nabla G(u),\ \text{and}\ \mathcal{H}(u;\zeta) = \frac{1}{2}(A - A^T)u + \zeta. \tag{99}$$

In this experiment, we generate the linear monotone operator A randomly by $A = B^T B + (C - C^T)$, where $B \in \mathbb{R}^{\lceil (n+1)/2\rceil \times (n+1)}$ (so that A is monotone but not strictly monotone), $C \in \mathbb{R}^{(n+1)\times(n+1)}$, and the entries of B and C are generated independently from the uniform $[0, 1]$ distribution. We generate ξ by $\xi \sim N(b, \frac{1}{n}I)$, where the entries of the mean vector b are also randomly distributed between 0 and 1. Therefore, in Assumption A1 we have $\sigma_G = 0$ and $\sigma_H = 1$. By setting $V(z,u) = \|z - u\|^2/2$, the prox-mapping $P_z^J(u)$ in (10) becomes the projection of $z - \eta$ to the Lorentz cone Z, which can be calculated efficiently. For the SAMP algorithm, we use the parameter settings in Corollary 2 with $L = \|A + A^T\|/2$, $M = \|A - A^T\|/2$, and $\beta = 1$. Since the study of the stochastic mirror-prox method in [20] only considers compact feasible sets, we could not find recommended parameter settings of the stochastic mirror-prox method for solving unbounded SVI. Therefore, we compare the SAMP algorithm with the deterministic mirror-prox (MP) method in [32]. In each iteration, the SAMP algorithm is supplied with the stochastic information $\mathcal{F}(u;\xi,\zeta)$, and the MP method is supplied with the deterministic information $F(u)$. We choose the parameters of MP according to (3.2) of [32] in which $L = \|A\|$. Noting the fact that SAMP computes relatively more matrix-vector multiplications due to the aforementioned decomposition, we set the total number of iterations of the MP method to be twice of that of the SAMP method. The performance of the SAMP and MP algorithms are compared in terms of the gap function (28), which is computed using MOSEK [30]. In particular, for any approximate solution w and perturbation vector v, we compute the value of $\tilde{g}(w,v)$ in (28) and norm of the perturbation vector $\|v\|$.[2] The comparison between the SAMP and MP algorithms is described in Table 4.

[2] See the proof of Theorem 2 for the definition of the perturbation term in the SAMP algorithm, and Theorem 5.2 in [28] for the definition of the perturbation term in the MP algorithm.

Two remarks on the performance of the SAMP and MP algorithms are in order. Firstly, it is interesting to observe that the practical convergence of the perturbation vector $\|v\|$ is slower than that of the gap function value $\tilde{g}(w, v)$, although they have the same rate of convergence (see Corollary 2). Secondly, the SAMP algorithm outperforms the MP algorithm for solving (98). Such performance comparison indicates the importance of the special treatment of the gradient field term ∇G from the SVI (98).

5 Concluding remarks

We present in this paper a novel stochastic accelerated mirror-prox method for solving a class of stochastic variational inequality problems. The basic idea of this algorithm is to incorporate a multi-step acceleration scheme into the stochastic mirror-prox method in [20]. The SAMP achieves the optimal iteration complexity, not only in terms of its dependence on the number of the iterations, but also on a variety of problem parameters. As a byproduct, the SAMP also significantly improves the iteration complexity for solving a class of deterministic variational inequalities. Moreover, the iteration cost of the SAMP is comparable to, or even less than that of the stochastic mirror-prox method in that it saves one computation of the stochastic gradient of the smooth component. To the best of our knowledge, this is the first algorithm with the optimal iteration complexity bounds for solving the SVIs of type (2). Furthermore, we show that the developed SAMP scheme can deal with the situation when the feasible region is unbounded, as long as a strong solution of the VI exists. In the unbounded case, we adopt the modified termination criterion employed by Monteiro and Svaiter in solving monotone inclusion problem, and demonstrate that the rate of convergence of SAMP depends on the distance from the initial point to the set of strong solutions. Our preliminary numerical results show that the proposed SAMP algorithm is promising to solve large-scale variational inequality problems.

It should be noted that in this paper we focus on the algorithm design for computing weak solutions to monotone SVIs. In view of some recent development on the unified analysis for convex and nonconvex stochastic optimization algorithms (see, e.g., [15]), it will be interesting to study a unified SAMP method that deals with both monotone and non-monotone SVIs. Also, considering that the problem of interest is a SVI with only deterministic feasible set, in the future it will be interesting to study different stochastic approximation type algorithms for solving SVIs involving expectation or probabilistic constraints. Moreover, it has been shown in [11,28,29] that the for $SVI(Z; 0, H, 0)$ with $\sigma = 0$ and $G = 0$, the mirror-prox method in [32] indeed converges to a strong solution with complexity $\mathcal{O}(M^2/\varepsilon^2)$. Since our main interest of this paper is to study the specialized treatment of gradient field ∇G in the SVI, we focus only on how to achieve the lower complexity bound for computing weak solutions. However, by incorporating the analysis in [11], it will be interesting to see if Algorithm 1 could also compute an approximate strong solution of the SVI problem (1) with complexity $\mathcal{O}(\sqrt{L/\varepsilon} + (M^2 + \sigma^2)/\varepsilon^2)$.

References

1. Auslender, A., Teboulle, M.: Interior projection-like methods for monotone variational inequalities. Math. Program. **104**, 39–68 (2005)
2. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**, 697–725 (2006)
3. Ben-Tal, A., Nemirovski, A.: Non-Euclidean restricted memory level method for large-scale convex optimization. Math. Program. **102**, 407–456 (2005)
4. Bennett, K.P., Mangasarian, O.L.: Robust linear programming discrimination of two linearly inseparable sets. Optim. Methods Softw. **1**, 23–34 (1992)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge university press, Cambridge (2004)
6. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR comput. Math. Math. Phys. **7**, 200–217 (1967)
7. Burachik, R.S., Iusem, A.N., Svaiter, B.F.: Enlargement of monotone operators with applications to variational inequalities. Set-Valued Anal. **5**, 159–180 (1997)
8. Chen, X., Wets, R.J.-B., Zhang, Y.: Stochastic variational inequalities: residual minimization smoothing sample average approximations. SIAM J. Optim. **22**, 649–673 (2012)
9. Chen, X., Ye, Y.: On homotopy-smoothing methods for box-constrained variational inequalities. SIAM J. Control Optim. **37**, 589–616 (1999)
10. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. SIAM J. Optim. **24**, 1779–1814 (2014)
11. Dang, C.D., Lan, G.: On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. Comput. Optim. Appl. **60**, 277–310 (2015)
12. Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, vol. 1. Springer, Berlin (2003)
13. Fang, S.C., Peterson, E.: Generalized variational inequalities. J. Optim. Theory Appl. **38**, 363–383 (1982)
14. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: a generic algorithmic framework. SIAM J. Optim. **22**, 1469–1492 (2012)
15. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**, 59–99 (2015)
16. Hartman, P., Stampacchia, G.: On some non-linear elliptic differential-functional equations. Acta Math. **115**, 271–310 (1966)
17. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Berlin (2009)
18. Jacob, L., Obozinski, G., Vert, J.-P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, pp. 433–440. (2009)
19. Jiang, H., Xu, H.: Stochastic approximation approaches to the stochastic variational inequality problem. IEEE Trans. Autom. Control **53**, 1462–1475 (2008)
20. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm. Stoch. Syst. **1**, 17–58 (2011)
21. Kien, B., Yao, J.-C., Yen, N.D.: On the solution existence of pseudomonotone variational inequalities. J. Glob. Optim. **41**, 135–145 (2008)
22. Korpelevich, G.: The extragradient method for finding saddle points and other problems. Matecon **12**, 747–756 (1976)
23. Koshal, J., Nedic, A., Shanbhag, U.V.: Regularized iterative stochastic approximation methods for stochastic variational inequality problems. IEEE Trans. Autom. Control **58**, 594–609 (2013)
24. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. **133**(1), 365–397 (2012)
25. Lan, G., Nemirovski, A., Shapiro, A.: Validation analysis of mirror descent stochastic approximation method. Math. Program. **134**, 425–458 (2012)
26. Lin, G.-H., Fukushima, M.: Stochastic equilibrium problems and stochastic mathematical programs with equilibrium constraints: a survey. Pac. J. Optim. **6**, 455–482 (2010)
27. Minty, G.J., et al.: Monotone (nonlinear) operators in hilbert space. Duke Math. J. **29**, 341–346 (1962)
28. Monteiro, R.D., Svaiter, B.F.: On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. SIAM J. Optim. **20**, 2755–2787 (2010)

29. Monteiro, R.D., Svaiter, B.F.: Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. SIAM J. Optim. **21**, 1688–1720 (2011)
30. MOSEK ApS, The MOSEK optimization toolbox for Matlab manual, version 6.0 (revision 135). MOSEK ApS, Denmark, (2012)
31. Nemirovski, A.: Information-based complexity of linear operator equations. J. Complex. **8**, 153–175 (1992)
32. Nemirovski, A.: Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex–concave saddle point problems. SIAM J. Optim. **15**, 229–251 (2004)
33. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**, 1574–1609 (2009)
34. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley-interscience series in discrete mathematics. Wiley, NewYork (1983)
35. Nesterov, Y.: Dual extrapolation and its applications to solving variational inequalities and related problems. Math. Program. **109**, 319–344 (2007)
36. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. **152**, 381–404 (2015)
37. Nesterov, Y., Vial, J.P.: Homogeneous analytic center cutting plane methods for convex problems and variational inequalities. SIAM J. Optim. **9**, 707–728 (1999)
38. Nesterov, Y.E.: A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Doklady SSSR **269**, 543–547 (1983). Translated as Soviet Math. Docl
39. Nesterov, Y.E.: Smooth minimization of nonsmooth functions. Math. Program. **103**, 127–152 (2005)
40. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)
41. Shapiro, A.: Monte carlo sampling methods. In: Shapiro, A., Ruszczyński, A. (eds.) Handbooks in Operations Research and Management Science, vol. 10, pp. 353–425. (2003)
42. Shapiro, A., Xu, H.: Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. Optimization **57**, 395–418 (2008)
43. Solodov, M.V., Svaiter, B.F.: A hybrid projection-proximal point algorithm. J. Convex Anal. **6**, 59–70 (1999)
44. Solodov, M.V., Svaiter, B.F.: An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. Math. Oper. Res. **25**, 214–230 (2000)
45. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B (Methodological), 267–288 (1996)
46. Tseng, P.: On accelerated proximal gradient methods for convex–concave optimization. Manuscript (2008)
47. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**, 25–57 (2006)
48. Wang, M., Lin, G., Gao, Y., Ali, M.M.: Sample average approximation method for a class of stochastic variational inequality problems. J. Syst. Sci. Complex. **24**, 1143–1153 (2011)
49. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. Adv. Neural Inf. Process. Syst. **15**, 505–512 (2003)
50. Xu, H., Zhang, D.: Stochastic nash equilibrium problems: sample average approximation and applications. Computa. Optim. Appl. **55**, 597–645 (2013)
51. Yousefian, F., Nedić, A., Shanbhag, U.V.: A regularized smoothing stochastic approximation (rssa) algorithm for stochastic variational inequality problems. In: Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World, IEEE Press, pp. 933–944. (2013)
52. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**, 301–320 (2005)