Noname manuscript No.
(will be inserted by the editor)

Accelerated Schemes for a Class of Variational Inequalities

Yunmei Chen · Guanghui Lan · Yuyuan Ouyang

Received: date / Accepted: date

Abstract We propose a class of novel methods, namely the accelerated mirror-prox (AMP) methods, for solving a class of deterministic and stochastic monotone variational inequalities (VI). The main idea of the proposed algorithms is to incorporate a multi-step acceleration scheme into the mirror-prox method. For both deterministic and stochastic VIs, the developed AMP methods compute the weak solutions with the optimal iteration complexity. In particular, if the monotone operator in VI consists of the gradient of a smooth function, the iteration complexities of the AMP methods can be accelerated in terms of their dependence on the Lipschitz constant of the smooth function. For VIs with bounded feasible sets, the bounds of the iteration complexities of the AMP methods depend on the diameter of the feasible set. For unbounded VIs, we adopt the modified gap function introduced by Monteiro and Svaiter for solving monotone inclusion, and show that the iteration complexities of the AMP methods depend on the distance from the initial point to the set of strong solutions. We also demonstrate the advantages of the AMP methods over some existing algorithms through our preliminary numerical experiments.

Keywords convex optimization · stochastic programming · variational inequalities · mirror-prox method · extragradient method

Yunmei Chen is partially supported by NSF grants DMS-1115568, IIP-1237814 and DMS-1319050. Guanghui Lan is partially supported by NSF grant CMMI-1000347, ONR grant N00014-13-1-0036, NSF DMS-1319050, and NSF CAREER Award CMMI-1254446. Part of the research was done while Yuyuan Ouyang was a PhD student at the Department of Mathematics, University of Florida, and Yuyuan Ouyang is partially supported by AFRL Mathematical Modeling Optimization Institute.

Yunmei Chen
Department of Mathematics, University of Florida
E-mail: yun@math.ufl.edu

Guanghui Lan
Department of Industrial and System Engineering, University of Florida
E-mail: glan@ise.ufl.edu

Yuyuan Ouyang
Department of Industrial and System Engineering, University of Florida
E-mail: ouyang@ufl.edu

Mathematics Subject Classification (2000) 90C25 · 90C15 · 62L20 · 68Q25

1 Introduction

Let $\mathcal{E}$ be a finite dimensional vector space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, and Z be a non-empty closed convex set in $\mathcal{E}$. Our problem of interest is to find $u^* \in Z$ that solves the following variational inequality (VI) problem:

$$\langle F(u), u^* - u \rangle \leq 0, \forall u \in Z, \tag{1} \quad \boxed{\texttt{eqnProblem}}$$

where F is defined by

$$F(u) = \nabla G(u) + H(u) + J'(u), \ \forall u \in Z. \tag{2} \quad \boxed{\texttt{eqnF}}$$

In (2), $G(\cdot)$ is a general continuously differentiable function whose gradient is Lipschitz continuous with constant L, i.e.,

$$0 \leq G(w) - G(v) - \langle \nabla G(w), w - v \rangle \leq \frac{L}{2} \|w - v\|^2, \forall w, v \in Z, \tag{3} \quad \boxed{\texttt{eqnGAssumption}}$$

$H : Z \to \mathcal{E}$ is a monotone operator with Lipschitz constant M, that is, for all $w, v \in Z$,

$$\langle H(w) - H(v), w - v \rangle \geq 0, \text{ and } \|H(w) - H(v)\|_* \leq M \|w - v\|, \tag{4} \quad \boxed{\texttt{eqnHAssumption}}$$

and $J'(u) \in \partial J(u)$, where $J(\cdot)$ is a relatively simple and convex function. We denote problem (1) by $VI(Z; G, H, J)$ or simply $VI(Z; F)$.

Observe that u^* given by (1) is often called a weak solution of $VI(Z; F)$. A related notion is a strong solution of VI. More specifically, we say that u^* is a strong solution of $VI(Z; F)$ if it satisfies

$$\langle F(u^*), u^* - u \rangle \leq 0, \forall u \in Z. \tag{5} \quad \boxed{\texttt{eqnSVI}}$$

For any monotone operator F, it is well-known that strong solutions of $VI(Z, F)$ are also weak solutions, and the reverse is also true under mild assumptions (e.g., when F is continuous). For example, for F in (2), if $J = 0$, then the weak and strong solutions of $VI(Z; G, H, 0)$ are equivalent.

The main goal of this paper is to develop efficient solution methods for solving two types of VIs, i.e., deterministic VIs with exact information about the operator F, and stochastic VIs where the operator F contains some stochastic components (e.g., ∇G and H) that cannot be evaluated exactly. We start by reviewing some existing methods for solving both these types of problems.

1.1 Deterministic VI

?⟨secIntroVID⟩? VI provides a unified framework for optimization, equilibrium and complementarity problems, and thus has been the focus of many algorithmic studies (see, e.g, [15,33, 8,32,36,38,23,30,20,14]). In particular, classical algorithms for VI include, but not limited to, the gradient projection method (e.g., [34,5]), Korpelevich's extragradient method [15], and the proximal point algorithm (e.g., [19,33]), etc. (see [11] for an extensive review and bibliography). While these earlier studies on VI solution methods focused on their asymptotic convergence behavior (see, e.g., [37,39,40]), much recent research effort has been devoted to algorithms exhibiting strong performance guarantees in a finite number of iterations (a.k.a., iteration complexity) [32,4,30,31,25,20,10]. More specifically, Nemirovski in a seminal work [23] presented a mirror-prox method by properly modifying Korpelevich's algorithm [16] and show that it can achieve an $\mathcal{O}(1/\epsilon)$ complexity bound for solving VI problems with Lipschitz continuous operators (i.e., smooth VI denoted by $VI(Z;0,H,0)$). Here $\epsilon > 0$ denotes the target accuracy in terms of a weak solution. This bound significantly improves the $\mathcal{O}(1/\epsilon^2)$ bound for solving VI problems with bounded operators (i.e., nonsmooth VI) (e.g., [4]). Nemirovski's algorithm was further generalized by Auslender and Teboulle [1] through the incorporation of a wider class of distance generating functions. Nesterov [30] has also developed a dual extrapolation method for solving smooth VI which possesses the same complexity bound as in [23]. More recently, Monteiro and Svaiter [20] showed that the hybrid proximal extragradient (HPE) method [35], which covers Korpelevich's algorithm as a special case, can also achieve the aforementioned $\mathcal{O}(1/\epsilon)$ complexity. Moreover, they developed novel termination criterion for VI problems with possibly unbounded feasible set Z, and derived the iteration complexity associated with HPE for solving unbounded VI problems accordingly. Monteiro and Svaiter [21] have also generalized the aforementioned $\mathcal{O}(1/\epsilon)$ complexity result for solving VI problems containing a simple nonsmooth component (i.e., $VI(Z;0,H,J)$).

It should be noted, however, that the aforementioned studies in the literature do not explore the fact that the operator F consists of a gradient component ∇G (see (2)). As a result, the iteration complexity associated with any of these algorithms, when applied to a smooth convex optimization problem (i.e., $VI(Z;G,0,0)$), is given by $\mathcal{O}(1/\epsilon)$, which is significantly worse than the well-known $\mathcal{O}(1/\sqrt{\epsilon})$ optimal complexity for smooth optimization [28]. An important motivating question for our study is whether one can utilize such structural properties of F in order to further improve the efficiency of VI solution methods. More specifically, we can easily see that the total number of gradient and operator evaluations for solving $VI(Z;G,H,J)$ cannot be smaller than

$$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}+\frac{M}{\epsilon}\right). \tag{6}$$

eqnOptRate

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

1. If $H=0$, $VI(Z;G,0,0)$ is equivalent to a smooth optimization problem $\min_{u\in Z} G(u)$, and the complexity for minimizing $G(u)$ cannot be better than $\mathcal{O}(\sqrt{L/\epsilon})$ [26,28];
2. If $G=0$, the complexity for solving $VI(Z;0,H,0)$ cannot be better than $\mathcal{O}(M/\epsilon)$ [27] (see also the discussions in Section 5 of [23]).

However, the best-known so-far iteration complexity bound for solving $VI(Z;G,H,J)$ is given by [14,20], where one needs to run these algorithms

$$\mathcal{O}\left(\frac{L+M}{\epsilon}\right), \tag{7}$$

eqnRateVI

iterations to compute a weak solution of $VI(Z;G,H,J)$, and each iteration requires the computation of both ∇G and H. It is worth noting that better iteration complexity bound has been achieved for a special case of $VI(Z;G,H,J)$ where the operator H is linear. In this case, Nesterov [29] showed that, by using a novel smoothing technique, the total number of first-order iterations (i.e., iterations requiring the computation of ∇G, the linear operators H and its conjugate H^*) for solving $VI(Z;G,H,J)$ can be bounded by (6). This bound has also been obtained by applying an accelerated primal-dual method recently developed by Chen, Lan and Ouyang [9]. Observe that the bound in (6) is significantly better than the one in (7) in terms of its dependence on L. However, it is unclear whether similar iteration complexity bounds to those in [29,9] can be achieved for the more general case when H is Lipschitz continuous.

1.2 Stochastic VI

While deterministic VIs had been intensively investigated in the literature, the study of stochastic VIs is still quite limited. In the stochastic setting, we assume that there exists *stochastic oracles* $\mathcal{SO}_G$ and $\mathcal{SO}_H$ that provide unbiased estimates to the operators $\nabla G(u)$ and $H(u)$ for any test point $u \in Z$. More specifically, we assume that 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 first-order information $\mathcal{G}(z, \xi_i)$ and $\mathcal{H}(z, \zeta_i)$ respectively, such that $\mathbb{E}[\mathcal{G}(x, \xi_i)] = \nabla G(x), \mathbb{E}[\mathcal{H}(x, \zeta_i)] = H(x)$, and

⟨itmVB⟩ **A1.** $\mathbb{E}\left[\|\mathcal{G}(x,\xi_i) - \nabla G(x)\|_*^2\right] \le \sigma_G^2,\ \mathbb{E}\left[\|\mathcal{H}(x,\zeta_i) - H(x)\|_*^2\right] \le \sigma_H^2,$

where $\xi_i \in \Xi$, $\zeta_i \in \Xi$ are independently distributed random variables. Throughout this paper, we may also denote

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

for the sake of notational convenience. It should be noted that deterministic VIs are special cases of stochastic VIs with $\sigma_G = \sigma_H = 0$. To distinguish stochastic VIs from their deterministic counterparts, we will use $SVI(Z; G, H, J)$ or simply $SVI(Z; F)$ to denote problem (1) under the aforementioned stochastic settings.

Following the discussion around (6) and the complexity theory for stochastic optimization [26,14], the total number of gradient and operator evaluations for solving stochastic VI cannot be smaller than

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

The best known complexity bound for computing $SVI(Z; G, H, 0)$ is given by the stochastic mirror-prox method in [14]. This method requires

$$\mathcal{O}\left(\frac{L+M}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right) \tag{10}$$ eqnBestRate

iterations to achieve the target accuracy $\epsilon > 0$ in terms of a weak solution, and each iteration requires the calls to $\mathcal{SO}_G$ and $\mathcal{SO}_H$. Similar to the deterministic case, the above complexity bound has been improved for some special cases, e.g., when $H = 0$ or H is linear. In particular, when $H = 0$, $SVI(Z, F)$ is equivalent to the stochastic minimization problem of $\min_{u \in Z} G(u) + J(u)$, Lan first presented in [17] (see more general results in [12,13]) an accelerated stochastic approximation method and showed that the iteration complexity of that algorithm is bounded by

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

More recently, Chen, Lan and Ouyang [9] presented an stochastic accelerated primal-dual method with a better complexity bound than (10) for solving $SVI(Z;G,H,J)$ with a linear operator H.

1.3 Contribution of this paper

Our contribution in this paper mainly consists of the following several aspects. Firstly, we present the accelerated mirror-prox (AMP) method for computing a solution of $VI(Z;G,H,J)$ by incorporating a multi-step acceleration scheme into the mirror-prox method in [23]. By utilizing the smoothness of $G(\cdot)$, we can significantly improve the iteration complexity from (7) to (6), while the iteration cost of AMP is comparable to that of the mirror-prox method. Therefore, AMP can solve VI problems efficiently with big Lipschitz constant L. To the best of our knowledge, this is the first time in the literature that such an optimal iteration complexity bound has been obtained for general Lipschitz continuous (rather than linear) operator H. We also present a simple backtracking strategy to estimate the proper choices of L and M.

Secondly, we develop a stochastic counterpart of AMP, namely SAMP, for solving $SVI(Z;G,H,J)$, and demonstrate that its iteration complexity for computing a weak solution is bounded by (9) and, similarly to the stochastic mirror-prox method, each iteration requires the calls to $\mathcal{SO}_G$ and $\mathcal{SO}_H$. Therefore, this algorithm improves the best-known complexity bounds for stochastic VI in terms of their dependence on the Lipschitz constant L. To the best of our knowledge, this is the first time that such an optimal iteration complexity bound has been developed for $SVI(Z;G,H,J)$ with general Lipschitz continuous (rather than linear) operator H. In addition, we investigate the stochastic VI method in more details, e.g., we develop the large-deviation results associated with the convergence of SAMP.

Thirdly, we incorporate into AMP the termination criterion employed by Monteiro and Svaiter [20,21] for solving variational and hemivariational inequalities posed as monotone inclusion problem. As a result, for both the deterministic and stochastic VIs, the AMP can deal with the case when Z is unbounded, as long as a strong solution to problem (5) exists, and the iteration complexity of AMP will depend on the distance from the initial point to the set of strong solutions.

Finally, we demonstrate the advantages of the developed AMP algorithms through preliminary numerical experiments on a few test problems.

1.4 Organization of the paper

The paper is organized as follows. We propose the AMP algorithms and discuss the main convergence results for solving deterministic VI and stochastic VI in Sections 2 and 3, respectively. To facilitate the readers, we present the proofs of the main convergence results in Section 4. Some preliminary numerical experiments are provided in Section 5 to demonstrate the efficiency of the AMP algorithms. Finally, we make some concluding remarks in Section 6.

2 Accelerated mirror-prox method for deterministic VI

⟨secAMP⟩ We introduce in this section an accelerated mirror-prox (AMP) method that computes a solution of $VI(Z; G, H, J)$, and discuss its main convergence properties.

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

$$P_z^J(\eta) := \operatorname*{argmin}_{u\in Z} \langle \eta, u - z\rangle + V(z,u) + J(u). \tag{11}$$ eqnProxMapping

In (11), the function $V(\cdot,\cdot)$ is defined by

$$V(z,u) := \omega(u) - \omega(z) - \langle \nabla\omega(z), u - z\rangle,\ \forall u, z \in Z, \tag{12}$$ eqnV

where $\omega(\cdot)$ is a strongly convex function with convexity parameter $\mu > 0$. The function $V(\cdot,\cdot)$ is known as a *prox-function*, or *Bregman divergence* [6] (see, e.g., [23,4,29,2] 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 AMP method in Algorithm 1.

Algorithm 1 The accelerated mirror-prox (AMP) method

```
⟨algAMP⟩
Choose $r_1 \in Z$. Set $w_1 = r_1$, $w_1^{ag} = r_1$.
For $t = 1, 2, \ldots, N-1$, calculate
    $w_t^{md} = (1-\alpha_t)w_t^{ag} + \alpha_t r_t,$   (13) eqnwmd
    $w_{t+1} = P_{r_t}^{\gamma_t J}\left(\gamma_t H(r_t) + \gamma_t \nabla G(w_t^{md})\right),$   (14) eqnProxwmd
    $r_{t+1} = P_{r_t}^{\gamma_t J}\left(\gamma_t H(w_{t+1}) + \gamma_t \nabla G(w_t^{md})\right),$   (15) eqnProxrmd
    $w_{t+1}^{ag} = (1-\alpha_t)w_t^{ag} + \alpha_t w_{t+1}.$   (16) eqnwag
Output $w_{N+1}^{ag}$.
```

Observe that the AMP method differs from the mirror-prox method in that we introduced two new sequences, i.e., $\{w_t^{md}\}$ and $\{w_t^{ag}\}$ (here "md" stands for "middle", and "ag" stands for "aggregated"). On the other hand, the mirror-prox method only has to compute the ergodic mean of the sequence $\{w_t\}$ as the output of the algorithm (similar to $\{w_t^{ag}\}$). If $\alpha_t \equiv 1$, $G = 0$ and $J = 0$, then Algorithm 1 for solving $VI(Z; 0, H, 0)$ is equivalent to the prox-method in [23]. In addition, if the distance generating function $w(\cdot) = \|\cdot\|^2/2$, then iterations (14) and (15) becomes

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

which are exactly the iterates of the extragradient method in [15]. On the other hand, if $H = 0$, then (14) and (15) produce the same optimizer $w_{t+1} = r_{t+1}$, and Algorithm 1 is equivalent 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 [41]). Therefore, Algorithm 1 can be viewed as a hybrid algorithm of the mirror-prox method and the accelerated gradient method, which gives its name accelerated mirror-prox method.

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

$$Q(\tilde{u}, u) := G(\tilde{u}) - G(u) + \langle H(u), \tilde{u} - u\rangle + J(\tilde{u}) - J(u). \qquad (17)\ \boxed{\texttt{eqnQ}}$$

Clearly, for F defined in (2), 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 $VI(Z; G, H, J)$. Hence when Z is bounded, it is natural to use the gap function

$$g(\tilde{u}) := \sup_{u\in Z} Q(\tilde{u}, u) \qquad (18)\ \boxed{\texttt{eqng0}}$$

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.

Theorem 1 below describes the convergence property 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:

$$\mathit{\Gamma}_t = \begin{cases} 1, & \text{when } t = 1 \\ (1-\alpha_t)\mathit{\Gamma}_{t-1}, & \text{when } t > 1, \end{cases} \qquad (19)\ \boxed{\texttt{eqnGamma}}$$

⟨thmAMPRateB⟩ **Theorem 1** *Suppose that*

$$\sup_{z_1,z_2\in Z} V(z_1, z_2) \leq \mathit{\Omega}_Z^2. \qquad (20)\ \boxed{\texttt{eqnVBounded}}$$

If the parameters $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *in Algorithm 1 are chosen such that* $\alpha_1 = 1$, *and*

$$0 \leq \alpha_t < 1,\ \ \mu - L\alpha_t\gamma_t - \frac{M^2\gamma_t^2}{\mu} \geq 0,\ \ \textit{and}\ \frac{\alpha_t}{\mathit{\Gamma}_t\gamma_t} \leq \frac{\alpha_{t+1}}{\mathit{\Gamma}_{t+1}\gamma_{t+1}},\ \forall t \geq 1, \qquad (21)\ \boxed{\texttt{eqnCondAlphaGammaBD}}$$

where $\{\mathit{\Gamma}_t\}$ *is defined by* (19), *and* μ *is the strong convexity parameter of* $\omega(\cdot)$ *in* (12). *Then,*

$$g(w_{t+1}^{ag}) \leq \frac{\alpha_t}{\gamma_t}\mathit{\Omega}_Z^2. \qquad (22)\ \boxed{\texttt{eqngBound}}$$

There are various options for choosing the parameters $\{\alpha_t\}$ and $\{\gamma_t\}$ that satisfy (21). In the following corollary, we give one example of such parameter settings.

⟨corAMPRateB⟩ **Corollary 1** *Suppose that* (20) *holds. If the parameters* $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *in AMP are set to*

$$\alpha_t = \frac{2}{t+1}\ \textit{and}\ \gamma_t = \frac{\mu t}{2(L + Mt)}, \qquad (23)\ \boxed{\texttt{eqnAlphaGammaBounded}}$$

then

$$g(w_{t+1}^{ag}) \leq \left(\frac{4L}{\mu t(t+1)} + \frac{4M}{\mu t}\right)\mathit{\Omega}_Z^2, \qquad (24)\ \boxed{\texttt{eqnAMPRateB}}$$

where $\mathit{\Omega}_Z$ *is defined in* (20).

Proof *Clearly,* $\Gamma_t = \dfrac{2}{t(t+1)}$ *satisfies* (19)*, and*

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

Moreover,

$$\mu - L\alpha_t\gamma_t - \frac{M^2\gamma_t^2}{\mu} = \mu - \frac{\mu L}{L+Mt}\cdot\frac{t}{t+1} - \frac{\mu M^2 t^2}{4(L+Mt)^2} \ge \mu - \frac{\mu L}{L+Mt} - \frac{\mu M t}{L+Mt} = 0.$$

Thus (21) *holds. Hence, by applying* (22) *in Theorem* 1 *with the parameter setting in* (23) *and using* (20)*, we obtain* (24)*.*

Clearly, in view of (24), when the parameters are chosen according to (23), 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. $g(\bar{w}) \le \epsilon$, can be bounded by

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

This bound significantly improves the best-known so-far complexity for solving problem (1) (see (6)) in terms of their dependence on the Lipschitz constant L. Moreover, it should be noted that the parameter setting in (23) is independent of Ω_Z, i.e., the AMP method achieves the above optimal iteration-complexity without requiring any information on the diameter of Z.

In Theorem 1, we assume that the Lipschitz constants L and M are known. In practice, the choices of these constants are critical, and wrong estimation for any of them may lead to slow convergence of the AMP method. However, by incorporating a simple backtracking strategy, we are able to search for the proper choices of L and M. The AMP algorithm with backtracking, as well as the theorem describing its convergence properties, are described in Algorithm 2 and Theorem 2.

⟨thmAMPRateBB⟩ **Theorem 2** *Suppose that* (20) *holds. Then the iterates* $\{w_{t+1}^{ag}\}$ *in Algorithm* 2 *satisfies*

$$g(w_{t+1}^{ag}) \le \left(\frac{4\max\{2L, L_0\}}{\mu t(t+1)} + \frac{4\max\{2M, M_0\}}{\mu t}\right)\Omega_Z^2,$$

where Ω_Z *is defined in* (20)*.*

A few remarks are in place for Algorithm 2. Firstly, if $L_0 = L$ and $M_0 = M$, then by the assumptions (3), (4) and the backtracking conditions (26) and (27), we obtain $L_t \equiv L$ and $M_t \equiv M$. In particular, Algorithm 2 reduces to Algorithm 1 in which the parameters are set to (23) in Corollary 1. Secondly, Algorithm 2 allows underestimation the Lipschitz constants L and M, without affecting its rate of convergence. Indeed, in view of the assumptions in (3), (4) and Steps 4 and 5, we can see that $L_t \le 2L$ and $M_t \le 2M$. Therefore, at any iteration t, the number of backtracking steps to search for L_t and M_t are less than $\lceil \log_2 2L/L_0 \rceil$ and $\lceil \log_2 2M/M_0 \rceil$, respectively.

Now, we consider the case when Z is unbounded. To study the convergence properties of AMP in this case, we use a perturbation-based termination criterion recently

Algorithm 2 The accelerated mirror-prox (AMP) method with backtracking

```
⟨algAMPB⟩ 1: Choose $r_1 \in Z, L_0 > 0$, and $M_0 > 0$. Set $w_1 = r_1$, $w_1^{ag} = r_1$, and $t = 1$.
⟨stepMain⟩ 2: Set $\hat{L}_t = L_{t-1}$ and $\hat{M}_t = M_{t-1}$.
⟨stepUpdate⟩ 3: Set the parameters to
        $$\alpha_t = \frac{2}{t+1} \text{ and } \hat{\gamma}_t = \frac{\mu t}{2(\hat{L}_t + \hat{M}_t t)}.$$   (25) eqngammat
⟨stepBTM⟩ 4: Compute $w_t^{md}$ and $w_{t+1}$ by equations (13) and (14) with $\gamma_t = \hat{\gamma}_t$. If
        $$\|H(w_{t+1}) - H(r_t)\|_* > \hat{M}_t \|w_{t+1} - r_t\|,$$   (26) eqnBTM
        then set $\hat{M}_t \leftarrow 2\hat{M}_t$, and go to Step 3.
⟨stepBTL⟩ 5: Compute $r_{t+1}$ and $w_{t+1}^{ag}$ by equations (15) and (16) with $\gamma_t = \hat{\gamma}_t$. If
        $$G(w_{t+1}^{ag}) - G(w_t^{md}) - \langle \nabla G(w_t^{md}), w_{t+1}^{ag} - w_t^{md} \rangle > \frac{\hat{L}_t}{2} \|w_{t+1}^{ag} - w_t^{md}\|^2,$$   (27) eqnBTL
        then set $\hat{L}_t \leftarrow 2\hat{L}_t$, and go to Step 3.
6: Set $L_t = \hat{L}_t$, $M_t = \hat{M}_t$, and $\gamma_t = \hat{\gamma}_t$.
7: If $t = N$, terminate and output $w_{N+1}^{ag}$. Otherwise, set $t \leftarrow t + 1$, and go to Step 2.
```

employed by Monteiro and Svaiter [20, 21], 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 $VI(Z; G, H, J)$ 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

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

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 3 below describes the convergence properties of AMP for solving deterministic VIs with unbounded feasible sets, under the assumption that a strong solution of (1) exists. It should be noted that this assumption does not limit too much the applicability of the AMP method. For example, when $J(\cdot) = 0$, any weak solution to $VI(Z; F)$ is also a strong solution.

⟨thmAMPRateUB⟩ **Theorem 3** *Suppose that $V(r, z) := \|z - r\|^2/2$ for any $r, z \in Z$. Also assume that the parameters $\{\alpha_t\}$ and $\{\gamma_t\}$ in Algorithm 1 are chosen such that $\alpha_1 = 1$, and for all $t > 1$,*

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

where Γ_t is defined in (19). Then for all $t \geq 1$ there exist $v_{t+1} = \alpha_t(r_1 - r_{t+1})/\gamma_t \in \mathcal{E}$ and $\varepsilon_{t+1} \geq 0$ such that $\tilde{g}(w_{t+1}^{ag}, v_{t+1}) \leq \varepsilon_{t+1}$. Moreover, we have

$$\|v_{t+1}\| \leq \frac{2\alpha_t D}{\gamma_t} \text{ and } \varepsilon_{t+1} \leq \frac{3\alpha_t(1 + \theta_t)D^2}{\gamma_t}. \tag{30}$$ eqnvepsThm

where

$$D := \|r_1 - u^*\|,\ \theta_t := \frac{\Gamma_t}{2(1-c^2)} \max_{i=1,\ldots,t} \frac{\alpha_i}{\gamma_i}. \qquad (31)\ \boxed{\texttt{eqndtheta}}$$

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

Below we provide a specific setting of parameters $\{\alpha_t\}$ and $\{\gamma_t\}$ that satisfies condition (29).

⟨corAMPRateUB⟩ **Corollary 2** *Suppose that* $V(r,z) := \|z-r\|^2/2$ *for any* $r, z \in Z$ *and* $M > 0$*. In Algorithm 1, if* $N \geq 2$ *is given and the parameters* $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *are set to*

$$\alpha_t = \frac{2}{t+1} \ \text{ and } \gamma_t = \frac{t}{3(L+MN)}, \qquad (32)\ \boxed{\texttt{eqnAlphaGammaUB}}$$

then there exists $v_N \in \mathcal{E}$ *such that* $\tilde{g}(w_N^{ag}, v_N) \leq \varepsilon_N$*,*

$$\|v_N\| \leq \left[\frac{12L}{N(N-1)} + \frac{12M}{N-1}\right] D, \ \text{ and } \varepsilon_N \leq \left[\frac{45L}{N(N-1)} + \frac{45M}{N-1}\right] D^2, \qquad (33)\ \boxed{\texttt{eqnvepsCor}}$$

where u^* *is a strong solution of* $VI(Z; F)$ *and* D *is defined in* (31)*.*

Proof *Clearly, we have* $\Gamma_t = 2/[t(t+1)]$ *and hence* (19) *is satisfied. It also follows from* (32) *that*

$$\begin{aligned} L\alpha_t\gamma_t + M^2\gamma_t^2 &= \frac{2Lt}{3(L+MN)(t+1)} + \frac{M^2t^2}{9(L+MN)^2} \leq \frac{2L}{3(L+MN)} + \frac{MN}{3(L+MN)} \\ &= \frac{2L+MN}{3(L+MN)} \leq \frac{2}{3} =: c^2. \end{aligned}$$

We can see that $c < 1$*, and* $1/(1-c^2) = 3$ *in* (31)*. Moreover, when* $N \geq 2$*,*

$$\theta_{N-1} = \frac{\Gamma_{N-1}}{2(1-c^2)} \max_{1\leq i\leq N-1} \{\frac{\alpha_i}{\Gamma_i}\} = \frac{1}{(1-c^2)N(N-1)} \max_{1\leq i\leq N-1} i = \frac{3}{N} \leq \frac{3}{2}.$$

We conclude (33) *by substituting the values of* α_{N-1}*,* γ_{N-1} *and* θ_{N-1} *to* (30)*.*

Several remarks are in place for the results obtained in Theorem 3 and Corollary 2. Firstly, although the existence of a strong solution u^* is assumed, no information on either u^* or D is needed for choosing parameters α_t and γ_t, as shown in (32) of Corollary 2. Secondly, both residuals $\|v_N\|$ and ε_N in (33) converge to 0 at the same rate (up to a constant $15D/4$). 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.

Algorithm 3 The stochastic accelerated mirror-prox (SAMP) method

⟨algAMPS⟩ Modify (14) and (15) in Algorithm 1 to

$$w_{t+1} = P_{r_t}^{\gamma_t J}\left(\gamma_t \mathcal{H}(r_t, \zeta_{2t-1}) + \gamma_t \mathcal{G}(w_t^{md}, \xi_t)\right), \tag{34}$$ eqnProxwmdS

$$r_{t+1} = P_{r_t}^{\gamma_t J}\left(\gamma_t \mathcal{H}(w_{t+1}, \zeta_{2t}) + \gamma_t \mathcal{G}(w_t^{md}, \xi_t)\right), \tag{35}$$ eqnProxrmdS

3 Accelerated mirror-prox method for stochastic VI

⟨secAMPS⟩ In this section, we focus on the $SVI(Z;F)$ and demonstrate that the stochastic counterpart of Algorithm 1 can achieve the optimal rate of convergence in (9).

The stochastic accelerated mirror-prox (SAMP) method is obtained by replacing the operators $H(r_t), H(w_{t+1})$ and $\nabla G(x_t^{md})$ in Algorithm 1 by their stochastic counterparts $\mathcal{H}(r_t, \zeta_{2t-1})$, $\mathcal{H}(w_{t+1}, \zeta_{2t})$ and $\mathcal{G}(w_t^{md}, \xi_t)$ respectively, by calling the stochastic oracles $\mathcal{SO}_G$ and $\mathcal{SO}_H$. This algorithm is formally described in Algorithm 3.

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 [14] 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 AMP is less than that of the stochastic mirror-prox method.

Similarly to Section 2, we use the gap function $g(\cdot)$ for the case when Z is bounded, and use the modified gap function $\tilde{g}(\cdot,\cdot)$ for the case when Z is unbounded. 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:

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

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

and

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

It should be noted that Assumption **A2.** implies Assumption **A1.** by Jensen's inequality.

The following theorem shows some convergence properties of Algorithm 3 when Z is bounded.

⟨thmAMPRateBS⟩ **Theorem 4** *Suppose that* (20) *holds. Also assume that the parameters* $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *in Algorithm 3 satisfy* $\alpha_1 = 1$,

$$q\mu - L\alpha_t\gamma_t - \frac{3M^2\gamma_t^2}{\mu} \ge 0 \text{ for some } q \in (0,1), \text{ and } \frac{\alpha_t}{\Gamma_t\gamma_t} \le \frac{\alpha_{t+1}}{\Gamma_{t+1}\gamma_{t+1}}, \ \forall t \ge 1, \tag{36}$$ eqnCondAlphaGammaIncS

where Γ_t *is defined in* (19). *Then,* *(a) Under Assumption* ***A1.****, for all* $t \ge 1$*,*

$$\mathbb{E}\left[g(w_{t+1}^{ag})\right] \le \mathcal{Q}_0(t) := \frac{2\alpha_t}{\gamma_t}\Omega_Z^2 + \left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right]\Gamma_t \sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}. \tag{37}$$

eqnQBoundBS

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

$$Prob\{g(w_{t+1}^{ag}) > \mathcal{Q}_0(t) + \lambda\mathcal{Q}_1(t)\} \le 2\exp\{-\lambda^2/3\} + 3\exp\{-\lambda\}, \tag{38}$$

eqnProb

where

$$\begin{aligned}\mathcal{Q}_1(t) := \Gamma_t(\sigma_G + \sigma_H)\Omega_Z\sqrt{\frac{2}{\mu}\sum_{i=1}^{t}\left(\frac{\alpha_i}{\Gamma_i}\right)^2} \\ + \left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right]\Gamma_t\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}.\end{aligned} \tag{39}$$

eqnQ1

We present below a specific parameter setting of $\{\alpha_t\}$ and $\{\gamma_t\}$ that satisfies (36).

⟨corStepS⟩ **Corollary 3** *Suppose that* (20) *holds. If the stepsizes* $\{\alpha_t\}$ *and* $\{\gamma_t\}$ *in Algorithm 3 are set to:*

$$\alpha_t = \frac{2}{t+1} \text{ and } \gamma_t = \frac{\mu t}{4L + 3Mt + \beta(t+1)\sqrt{\mu t}}, \tag{40}$$

eqnAlphaGammaBDS

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

$$\mathbb{E}\left[g(w_{t+1}^{ag})\right] \le \frac{16L\Omega_Z^2}{\mu t(t+1)} + \frac{12M\Omega_Z^2}{\mu(t+1)} + \frac{\sigma\Omega_Z}{\sqrt{\mu(t-1)}}\left(\frac{4\beta\Omega_Z}{\sigma} + \frac{16\sigma}{3\beta\Omega_Z}\right) =: \mathcal{C}_0(t), \tag{41}$$

eqnC0

where σ *and* Ω_Z *are defined in* (8) *and* (20)*, respectively. Furthermore, under Assumption* ***A2.****,*

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

where

$$\mathcal{C}_1(t) := \frac{\sigma\Omega_Z}{\sqrt{\mu(t-1)}}\left(\frac{4\sqrt{3}}{3} + \frac{16\sigma}{3\beta\Omega_Z}\right). \tag{42}$$

eqnC1

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 (40), 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 (36) *holds with constant* $q = 5/6$. *In view of Theorem 4, 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*

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

Using the above inequality, (20), (37), (39), (40), *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

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

In view of (9), (41) and (42), 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 addition, we can also see that this algorithm allows L to be as large as $\mathcal{O}(t^{3/2})$ without significantly affecting its convergence properties.

In the following theorem, we demonstrate some convergence properties of Algorithm 3 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.

⟨thmAMPRateUBS⟩ **Theorem 5** *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* $t > 1$,

$$0 \le \alpha_t < 1,\ L\alpha_t\gamma_t + 3M^2\gamma_t^2 \le c^2 < q \text{ for some } c, q \in (0,1), \text{ and } \frac{\alpha_t}{\Gamma_t \gamma_t} = \frac{\alpha_{t+1}}{\Gamma_{t+1}\gamma_{t+1}}, \tag{43}$$

eqnCondAlphaGammaUBS

where Γ_t *is defined in* (19). *Then for all* $t \ge 1$ *there exists a perturbation vector* v_{t+1} *and a residual* $\varepsilon_{t+1} \ge 0$ *such that* $\tilde{g}(w_{t+1}^{ag}, v_{t+1}) \le \varepsilon_{t+1}$. *Moreover, for all* $t \ge 1$, *we have*

$$\mathbb{E}[\|v_{t+1}\|] \le \frac{\alpha_t}{\gamma_t} \left(2D + 2\sqrt{D^2 + C_t^2}\right), \tag{44}$$

eqnEv

$$\mathbb{E}[\varepsilon_{t+1}] \le \frac{\alpha_t}{\gamma_t} \left[(3 + 6\theta)D^2 + (1 + 6\theta)C_t^2\right] + \frac{18\alpha_t^2\sigma_H^2}{\gamma_t^2} \sum_{i=1}^{t} \gamma_i^3, \tag{45}$$

eqnEeps

where u^* is a strong solution of $VI(Z; G, H, J)$, D is defined in (31),

$$\theta = \max\left\{1, \frac{c^2}{q-c^2}\right\} \text{ and } C_t = \sqrt{\left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right]\sum_{i=1}^{t}\gamma_i^2}. \qquad \boxed{\texttt{eqnCtheta}} \tag{46}$$

Below we give an example of parameters α_t and γ_t that satisfies (43).

?⟨corStepUBS⟩? **Corollary 4** *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 3 are set to*

$$\alpha_t = \frac{2}{t+1} \text{ and } \gamma_t = \frac{t}{5L + 3MN + \beta N\sqrt{N-1}}, \qquad \boxed{\texttt{eqnStepUBS}} \tag{47}$$

where σ is defined in Corollary 3, then there exists $v_N \in \mathcal{E}$ and $\varepsilon_N > 0$, such that $\tilde{g}(w_N^{ag}, v_N) \le \varepsilon_N$,

$$\mathbb{E}[\|v_N\|] \le \frac{40LD}{N(N-1)} + \frac{24MD}{N-1} + \frac{\sigma}{\sqrt{N-1}}\left(\frac{8\beta D}{\sigma} + 5\right), \qquad \boxed{\texttt{eqnEvUB}} \tag{48}$$

and

$$\mathbb{E}[\varepsilon_N] \le \frac{90LD^2}{N(N-1)} + \frac{54MD^2}{N-1} + \frac{\sigma D}{\sqrt{N-1}}\left(\frac{18\beta D}{\sigma} + \frac{56\sigma}{3\beta D} + \frac{18\sigma}{\beta DN}\right). \qquad \boxed{\texttt{eqnEepsUB}} \tag{49}$$

Proof Clearly, we have $\Gamma_t = 2/[t(t+1)]$, and hence (19) is satisfied. Moreover, in view of (47), we have

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

which implies that (43) is satisfied with $c^2 = 5/12$ and $q = 5/6$. Observing from (47) that $\gamma_t = t\gamma_1$, setting $t = N-1$ in (46) and (47), we obtain

$$\frac{\alpha_{N-1}}{\gamma_{N-1}} = \frac{2}{\gamma_1 N(N-1)} \text{ and } C_{N-1}^2 = 4\sigma^2\sum_{i=1}^{N-1}\gamma_1^2 i^2 \le \frac{4\sigma^2\gamma_1^2N^2(N-1)}{3}, \qquad \boxed{\texttt{eqnCN}} \tag{50}$$

where C_{N-1} is defined in (46). Applying (50) to (44) 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 (45), (50), and the facts that $\theta = 1$ in (46) and

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

we have

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

Observe that we need to choose a parameter β for the stochastic unbounded case, which is not required for the deterministic case (see Corollary 2). One may want to choose β in a way such that the right hand side of (48) or (49) 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 (48) and (49) 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.

4 Convergence analysis

⟨secProof⟩ In this section, we focus on proving the main convergence results in Sections 2 and 3, namely, Theorems 1, 3, 4 and 5.

4.1 Convergence analysis for deterministic AMP

?⟨secProofD⟩? In this section, we prove Theorems 1 and 3 in Section 2, which state the main convergence properties of Algorithm 1 for solving the deterministic problem $VI(Z;G,H,J)$.

To prove the convergence of the deterministic 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 (14) and (15) of Algorithm 1. Lemma 3 provides a recursion related to the function $Q(\cdot,\cdot)$ defined in (17). With the help of Lemmas 1, 2 and 3, we estimate a bound on $Q(\cdot,\cdot)$ in Proposition 1.

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

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

Proof *See Lemma 2 in [12] for the proof.*

The following proposition is a slight extension of Lemma 6.3 in [14]. In particular, when $J(\cdot) = 0$, we can obtain (54) and (55) directly by applying (53) to (6.8) in [14], and the results when $J(\cdot) \not\equiv 0$ can be easily constructed from the proof of Lemma 6.3 in [14]. We provide the proof here only for the sake of completeness.

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

$$w = P_r^J(\eta), \tag{51}$$ eqnProx1

$$y = P_r^J(\vartheta), \tag{52}$$ eqnProx2

and

$$\|\vartheta - \eta\|_*^2 \le L^2\|w - r\|^2 + M^2. \tag{53}$$ eqnLM

Then, for all $u \in Z$,

$$\langle \vartheta, w - u\rangle + J(w) - J(u) \le V(r, u) - V(y, u) - \left(\frac{\mu}{2} - \frac{L^2}{2\mu}\right)\|r - w\|^2 + \frac{M^2}{2\mu}, \tag{54}$$ eqnPRecursion

and

$$V(y, w) \le \frac{L^2}{\mu^2}V(r, w) + \frac{M^2}{2\mu}. \tag{55}$$ eqnVvr

Proof *Applying Proposition 1 to (51) and (52), for all $u \in Z$ we have*

$$\langle \eta, w - u\rangle + J(w) - J(u) \le V(r, u) - V(r, w) - V(w, u), \tag{56}$$ eqnMP2u

$$\langle \vartheta, y - u\rangle + J(y) - J(u) \le V(r, u) - V(r, y) - V(y, u), \tag{57}$$ eqnMP1

In particular, letting $u = y$ in (56) we have

$$\langle \eta, w - y\rangle + J(w) - J(y) \le V(r, y) - V(r, w) - V(w, y). \tag{58}$$ eqnMP2

Adding inequalities (57) and (58), then

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

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

$$\frac{\mu}{2}\|z - u\|^2 \le V(u, z), \forall u, z \in Z, \tag{59}$$ eqnVvsNorm

due to the strong convexity of $\omega(\cdot)$ in (12), we obtain

$$\begin{aligned}
&\langle \vartheta, w - u\rangle + J(w) - J(u) \\
\le\ & \|\vartheta - \eta\|_*\|w - y\| + V(r, u) - V(y, u) - V(r, w) - \frac{\mu}{2}\|w - y\|^2 \\
\le\ & \frac{1}{2\mu}\|\vartheta - \eta\|_*^2 + \frac{\mu}{2}\|w - y\|^2 + V(r, u) - V(y, u) - V(r, w) - \frac{\mu}{2}\|w - y\|^2 \qquad (60) \\
=\ & \frac{1}{2\mu}\|\vartheta - \eta\|_*^2 + V(r, u) - V(y, u) - V(r, w).
\end{aligned}$$ tmp1

The result in (54) then follows immediately from above relation, (53) and (59).

Moreover, observe that by setting $u = w$ and $u = y$ in (57) and (60), respectively, we have

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

Adding the above two inequalities, and using (53) and (59), we have

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

and thus (55) holds.

⟨propSimplifiedQ⟩ **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 (13) and (16), then for all $u \in Z$,*

$$\begin{aligned} & Q(w_{t+1}^{ag}, u) - (1 - \alpha_t)Q(w_t^{ag}, u) \\ & \le \alpha_t\langle \nabla G(w_t^{md}) + H(w_{t+1}), w_{t+1} - u\rangle + \frac{L\alpha_t^2}{2}\|w_{t+1} - r_t\|^2 - \alpha_t J(u). \end{aligned} \tag{61}$$ eqnSimplifiedQ

Proof *Observe from (13) and (16) 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 $u \in Z$,*

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

Using the above inequality, (16), (17) 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 Lemma 3, we assume that the Lipschitz constant L satisfies (3). It can be easily seen that for any $L_t > 0$, as long as

$$G(w_{t+1}^{ag}) \le G(w_t^{md}) + \langle \nabla G(w_t^{md}), w_{t+1}^{ag} - w_t^{md} \rangle + \frac{L_t}{2} \|w_{t+1}^{ag} - w_t^{md}\|^2, \tag{62}$$

eqnGLt

then the above lemma still holds with L in (61) replaced by L_t:

$$\begin{aligned} & Q(w_{t+1}^{ag}, u) - (1-\alpha_t) Q(w_t^{ag}, u) \\ & \le \alpha_t \langle \nabla G(w_t^{md}) + H(w_{t+1}), w_{t+1} - u \rangle + \frac{L_t \alpha_t^2}{2} \|w_{t+1} - r_t\|^2 - \alpha_t J(u). \end{aligned} \tag{63}$$

?eqnSimplifiedQBT?

The following proposition estimates a bound on $Q(w_{t+1}^{ag}, u)$, and will be used in the proof of both Theorems 1 and 3.

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

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(w_{t+1}^{ag}, u) \\ & \le \mathcal{B}_t(u, r_{[t]}) - \sum_{i=1}^{t} \frac{\alpha_i}{2\Gamma_i \gamma_i} \left(\mu - L\alpha_i \gamma_i - \frac{M^2 \gamma_i^2}{\mu} \right) \|r_i - w_{i+1}\|^2, \ \forall u \in Z, \end{aligned} \tag{64}$$

eqnQBoundGeneral

where Γ_t is defined in (19), and

$$\mathcal{B}_t(u, r_{[t]}) := \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i \gamma_i} (V(r_i, u) - V(r_{i+1}, u)). \tag{65}$$

eqnB

Proof *First, it follows from Lemma 2 applied to iterations (14) and (15) (with $r = r_t, w = w_{t+1}, y = r_{t+1}, \vartheta = \gamma_t H(r_t) + \gamma_t \nabla G(w_t^{md}), \eta = \gamma_t H(w_{t+1}) + \gamma_t \nabla G(w_t^{md}), J = \gamma_t J$, $L = M\gamma_t$ and $\nu = 0$) that for any $u \in Z$,*

$$\begin{aligned} & \gamma_t \langle \nabla G(w_t^{md}) + H(w_{t+1}), w_{t+1} - u \rangle + \gamma_t J(w_{t+1}) - \gamma_t J(u) \\ & \le V(r_t, u) - V(r_{t+1}, u) - \left(\frac{\mu}{2} - \frac{M^2 \gamma_t^2}{2\mu} \right) \|r_t - w_{t+1}\|^2. \end{aligned}$$

Now applying the above inequality to (61), we have

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

Dividing both sides of the above inequality by Γ_t, we have

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

Using the facts that $\alpha_1 = 1$, and that $\dfrac{1-\alpha_t}{\Gamma_t} = \dfrac{1}{\Gamma_{t-1}}$, $t > 1$, due to (19), we can apply the above inequality recursively to obtain (64).

It is not difficult to see from the proof of the above lemma that it can be slightly modified for variable Lipschitz constants L_t and M_t. Indeed, for any series $\{L_t\}$ and $\{M_t\}$ that satisfy (62) and

$$\|H(w_{t+1}) - H(r_t)\|_* \leq M_t \|w_{t+1} - r_t\|, \tag{66}$$

eqnHMt

respectively, we have

$$\begin{aligned} &\frac{1}{\Gamma_t} Q(w^{ag}_{t+1}, u) \\ &\leq \mathcal{B}_t(u, r_{[t]}) - \sum_{i=1}^{t} \frac{\alpha_i}{2\Gamma_i \gamma_i}\left(\mu - L_i\alpha_i\gamma_i - \frac{M_i^2\gamma_i^2}{\mu}\right)\|r_i - w_{i+1}\|^2, \ \forall u \in Z. \end{aligned} \tag{67}$$

eqnQBoundGeneralBT

We will use equation (67) in the proof of Theorem 2 for the convergence of the AMP algorithm with backtracking.

We are now ready to prove Theorem 1, which provides an estimate of the gap function of the deterministic AMP algorithm when Z is bounded. This result follows immediately from Lemma 1.

Proof of Theorem 1. In view of (21) *and* (64)*, to prove* (22) *it suffices to show that* $\mathcal{B}_t(u, r_{[t]}) \leq \alpha_t \Omega_Z^2/(\Gamma_t\gamma_t)$ *for all* $u \in Z$*. Indeed, since the sequence* $\{r_i\}_{i=1}^{t+1}$ *is in the bounded set* Z*, applying* (20) *and* (21) *to* (65) *we have*

$$\begin{aligned} &\mathcal{B}_t(u, r_{[t]}) \\ &= \frac{\alpha_1}{\Gamma_1\gamma_1} V(r_1, u) - \sum_{i=1}^{t-1}\left[\frac{\alpha_i}{\Gamma_i\gamma_i} - \frac{\alpha_{i+1}}{\Gamma_{i+1}\gamma_{i+1}}\right] V(r_{i+1}, u) - \frac{\alpha_t}{\Gamma_t\gamma_t} V(r_{t+1}, u) \\ &\leq \frac{\alpha_1}{\Gamma_1\gamma_1}\Omega_Z^2 - \sum_{i=1}^{t-1}\left[\frac{\alpha_i}{\Gamma_i\gamma_i} - \frac{\alpha_{i+1}}{\Gamma_{i+1}\gamma_{i+1}}\right]\Omega_Z^2 = \frac{\alpha_t}{\Gamma_t\gamma_t}\Omega_Z^2, \ \forall u \in Z, \end{aligned} \tag{68}$$

eqnBBD

and thus (22) *holds.*

Similar as the proof of the above theorem, with the help of Lemma 1, (62) and (66), we are able to prove Theorem 2.

Proof of Theorem 2. We have $\Gamma_t = 2/[t(t+1)]$ *and hence* (19) *is satisfied. Also, in view of the remarks after Algorithm 2, we have* $L_t \leq \min\{2L, L_0\}$ *and* $M_t \leq \min\{2M, M_0\}$*. Moreover, from Steps 2, 4 and 5 of Algorithm 2 we can see that the sequences* $\{L_t\}_{t\geq 0}$ *and* $\{M_t\}_{t\geq 0}$ *are non-decreasing. Following these observations together with* (25)*, we have*

$$\frac{\alpha_t}{\Gamma_t\gamma_t} = \frac{2}{\mu}(L_t + M_t t) \leq \frac{\alpha_{t+1}}{\Gamma_{t+1}\gamma_{t+1}},$$

and

$$\begin{aligned} &\mu - L_t\alpha_t\gamma_t - \frac{M_t^2\gamma_t^2}{\mu} = \mu - \frac{\mu L_t}{L_t + M_t t}\cdot\frac{t}{t+1} - \frac{\mu M_t^2 t^2}{4(L_t + M_t t)^2} \\ &\geq \mu - \frac{\mu L_t}{L_t + M_t t} - \frac{\mu M_t t}{L_t + M_t t} = 0. \end{aligned}$$

Applying the above two inequalities, (25), and (68) to (64), we conclude that

$$Q(w_{t+1}^{ag}, u) \le \Gamma_t \mathcal{B}_t(u, r_{[t]}) \le \frac{\alpha_t}{\gamma_t}\Omega_Z^2 = \left(\frac{4L_t}{\mu t(t+1)} + \frac{4M_t}{\mu(t+1)}\right)\Omega_Z^2, \ \forall u \in Z.$$

In the remaining part of this subsection, we will focus on proving Theorem 3, which summarizes some convergence properties of the deterministic AMP algorithm when Z is unbounded.

Proof of Theorem 3. Using the assumption that $V(r,z) := \|z-r\|^2/2$ *for all* $r, z \in Z$, *and applying the last relation of (29) to (65), we obtain*

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

Applying the above identity and the second relation of (29) to (64) and noting that $\mu = 1$, *we have*

$$Q(w_{t+1}^{ag}, u) \le \frac{\alpha_t}{2\gamma_t}\|r_1 - u\|^2 - \frac{\alpha_t}{2\gamma_t}\|r_{t+1} - u\|^2 - \frac{\alpha_t}{2\gamma_t}\sum_{i=1}^{t}\left(1 - c^2\right)\|r_i - w_{i+1}\|^2. \qquad (69)\ \boxed{\texttt{eqnQEucl}}$$

Observing that

$$\begin{aligned}&\frac{1}{2}\|r_1 - u\|^2 - \frac{1}{2}\|r_{t+1} - u\|^2 = \frac{1}{2}\|r_1\|^2 - \frac{1}{2}\|r_{t+1}\|^2 - \langle r_1 - r_{t+1}, u\rangle\\ &= \frac{1}{2}\|r_1 - w_{t+1}^{ag}\|^2 - \frac{1}{2}\|r_{t+1} - w_{t+1}^{ag}\|^2 + \langle r_1 - r_{t+1}, w_{t+1}^{ag} - u\rangle,\end{aligned} \qquad (70)\ \boxed{\texttt{eqnr2wag}}$$

and combining (69) and (70), we obtain

$$\begin{aligned}&Q(w_{t+1}^{ag}, u) - \frac{\alpha_t}{\gamma_t}\langle r_1 - r_{t+1}, w_{t+1}^{ag} - u\rangle\\ &\le \frac{\alpha_t}{2\gamma_t}\|r_1 - w_{t+1}^{ag}\|^2 - \frac{\alpha_t}{2\gamma_t}\|r_{t+1} - w_{t+1}^{ag}\|^2 - \frac{\alpha_t}{2\gamma_t}(1-c^2)\sum_{i=1}^{t}\|r_i - w_{i+1}\|^2 =: \varepsilon_{t+1}.\end{aligned} \qquad (71)\ \boxed{\texttt{eqneps}}$$

Therefore, if we set $v_{t+1} := \alpha_t(r_1 - r_{t+1})/\gamma_t$, *then* $Q(w_{t+1}^{ag}, u) - \langle v_{t+1}, w_{t+1}^{ag} - u\rangle \le \varepsilon_{t+1}$ *for all* $u \in Z$. *Note that* $\varepsilon_{t+1} \ge 0$ *holds trivially by letting* $u = w_{t+1}^{ag}$ *in (71). Hence we have* $\tilde{g}(w_{t+1}^{ag}, v_{t+1}) \le \varepsilon_{t+1}$ *and it suffices to estimate the bounds on* $\|v_{t+1}\|$ *and* ε_{t+1}.

Observe that by (2), (5), (17) and the convexity of G *and* J, *we have*

$$Q(w_{t+1}^{ag}, u^*) \ge \langle \nabla F(u^*), w_{t+1}^{ag} - u^*\rangle \ge 0, \qquad (72)\ \boxed{\texttt{eqnQpositive}}$$

where the last inequality follows from the assumption that u^* *is a strong solution of* $VI(Z; G, H, J)$. *This observation together with (69) imply that*

$$\|r_1 - u^*\|^2 - \|r_{t+1} - u^*\|^2 - \sum_{i=1}^{t}\left(1 - c^2\right)\|r_i - w_{i+1}\|^2 \ge 0.$$

By the above inequality and the definition of D in (31), we have

$$\begin{aligned}\|r_{t+1}-u^*\| \leq & D, \\ \sum_{i=1}^{t}\|r_i - w_{i+1}\|^2 \leq & \frac{D^2}{1-c^2}.\end{aligned}$$

(73) eqnrtnBound

(74) eqnrtwtnBound

It then follows from (73) and the definition of v_{t+1} that

$$\|v_{t+1}\| \leq \frac{\alpha_t}{\gamma_t}\left(\|r_1-u^*\|+\|r_{t+1}-u^*\|\right) \leq \frac{2\alpha_t}{\gamma_t}D,$$

and hence the first relation in (30) holds.

To finish the proof, it now suffices to estimate a bound for ε_t. Firstly we explore the definition of the aggregate point w_{t+1}^{ag}. By (16) and (19), we have

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

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

$$w_{t+1}^{ag} = \Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}w_{i+1}, \tag{75}$$

eqnwagReform

where by (19) we have

$$\Gamma_t \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i} = 1. \tag{76}$$

eqnGammaSpan

Therefore, w_{t+1}^{ag} is a convex combination of iterates $w_2, \ldots, w_{t+1}$. Using (31), (71), (73) and (74), we conclude that

$$\begin{aligned}\varepsilon_{t+1} \leq & \frac{\alpha_t}{2\gamma_t}\|r_1 - w_{t+1}^{ag}\|^2 \leq \frac{\alpha_t \Gamma_t}{2\gamma_t}\sum_{i=1}^{t}\frac{\alpha_i}{\gamma_i}\|r_1 - w_{i+1}\|^2 \\ \leq & \frac{3\alpha_t\Gamma_t}{2\gamma_t}\sum_{i=1}^{t}\frac{\alpha_i}{\gamma_i}(\|r_1-u^*\|^2 + \|r_i - u^*\|^2 + \|r_i - w_{i+1}\|^2) \\ \leq & \frac{3\alpha_t}{2\gamma_t}\left(2D^2 + \Gamma_t \max_{i=1,\ldots,t}\frac{\alpha_i}{\gamma_i}\sum_{i=1}^{t}\|r_i - w_{i+1}\|^2\right) \\ \leq & \frac{3\alpha_t(1+\theta_t)D^2}{\gamma_t}.\end{aligned}$$

4.2 Convergence analysis for stochastic AMP

?⟨secProofS⟩? In this section, we prove the convergence results of the SAMP method presented in Section 3, namely, Theorems 4 and 5.

Throughout this section, 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

$$\begin{aligned} \Delta_H^{2t-1} &:= \mathcal{H}(r_t, \zeta_{2t-1}) - H(r_t), \\ \Delta_H^{2t} &:= \mathcal{H}(w_{t+1}, \zeta_{2t}) - H(w_{t+1}), \text{ and} \\ \Delta_G^{t} &:= \mathcal{G}(w_t^{md}, \xi_t) - \nabla G(w_t^{md}). \end{aligned} \tag{77}$$

eqnDelta

To start with, we present a technical result to obtain a bound on $Q(w_{t+1}^{ag}, u)$ for all $u \in Z$. The following lemma is analogous to Lemma 1 for deterministic AMP, and will be applied in the proof of Theorems 4 and 5.

roQBoundGeneralS⟩ **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 $\{w_t^{ag}\}$ generated by Algorithm 3 satisfy*

$$\begin{aligned} &\frac{1}{\Gamma_t} Q(w_{t+1}^{ag}, u) \\ &\le \mathcal{B}_t(u, r_{[t]}) - \sum_{i=1}^{t} \frac{\alpha_i}{2\Gamma_i \gamma_i}\left(q\mu - L\alpha_i\gamma_i - \frac{3M^2\gamma_i^2}{\mu}\right)\|r_i - w_{i+1}\|^2 + \sum_{i=1}^{t} \Lambda_i(u), \ \forall u \in Z, \end{aligned} \tag{78}$$

eqnQBoundGeneralS

where Γ_t is defined in (19), $\mathcal{B}_t(u, r_{[t]})$ is defined in (65), and

$$\begin{aligned} \Lambda_i(u) := &\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2\right) - \frac{(1-q)\mu\alpha_i}{2\Gamma_i\gamma_i}\|r_i - w_{i+1}\|^2 \\ &- \frac{\alpha_i}{\Gamma_i}\langle \Delta_H^{2i} + \Delta_G^{i}, w_{i+1} - u\rangle. \end{aligned} \tag{79}$$

eqnLambda

Proof *Observe from (77) that*

$$\begin{aligned} &\|\mathcal{H}(w_{t+1}, \zeta_{2t}) - \mathcal{H}(r_t, \zeta_{2t-1})\|_*^2 \\ \le\ &\left(\|H(w_{t+1}) - H(r_t)\|_* + \|\Delta_H^{2t}\|_* + \|\Delta_H^{2t-1}\|_*\right)^2 \\ \le\ &3\left(\|H(w_{t+1}) - H(r_t)\|_*^2 + \|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2\right) \\ \le\ &3\left(M^2\|w_{t+1} - r_t\|^2 + \|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2\right). \end{aligned} \tag{80}$$

eqnLMStoc

Applying Proposition 2 to (34) and (35) (with $r = r_t, w = w_{t+1}, y = r_{t+1}, \vartheta = \gamma_t\mathcal{H}(r_t, \zeta_{2t-1}) + \gamma_t\mathcal{G}(w_t^{md}, \xi_t)$, $\eta = \gamma_t\mathcal{H}(w_{t+1}, \zeta_{2t}) + \gamma_t\mathcal{G}(w_t^{md}, \xi_t), J = \gamma_t J$, $L^2 = 3M^2\gamma_t^2$ and $M^2 = 3\gamma_t^2(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2)$), and using (80), we have for any $u \in Z$,

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

Applying (77) *and the above inequality to* (61)*, we have*

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

Dividing the above inequality by $\varGamma_t$ *and using the definition of* $\varLambda_t(u)$ *in* (79)*, we obtain*

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

Noting the fact that $\alpha_1=1$ *and* $(1-\alpha_t)/\varGamma_t=1/\varGamma_{t-1}$*,* $t>1$*, due to* (19)*, applying the above inequality recursively and using the definition of* $\mathcal{B}_t(\cdot,\cdot)$ *in* (65)*, we conclude* (78)*.*

We still need the following technical result to prove Theorem 4.

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

$$w_{i+1}^v=\operatorname*{argmin}_{u\in Z}-\gamma_i\langle\varDelta^i,u\rangle+V(w_i^v,u),\ \forall i>1, \tag{81}$$ eqnProxv

then

$$\sum_{i=1}^t\theta_i\langle-\varDelta^i,w_i^v-u\rangle\le\sum_{i=1}^t\frac{\theta_i}{\gamma_i}(V(w_i^v,u)-V(w_{i+1}^v,u))+\sum_{i=1}^t\frac{\theta_i\gamma_i}{2\mu}\|\varDelta_i\|_*^2,\ \forall u\in Z. \tag{82}$$ eqnTech

Proof *Applying Lemma 1 to* (81) *(with* $r=w_i^v$*,* $w=w_{i+1}^v$*,* $\zeta=-\gamma_i\varDelta^i$ *and* $J=0$*), we have*

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

Moreover, by Schwartz inequality, Young's inequality and (59) *we have*

$$\begin{aligned}
&-\gamma_i\langle\varDelta^i,w_i^v-w_{i+1}^v\rangle\\
&\le\gamma_i\|\varDelta^i\|_*\|w_i^v-w_{i+1}^v\|\le\frac{\gamma_i^2}{2\mu}\|\varDelta_i\|_*^2+\frac{\mu}{2}\|w_i^v-w_{i+1}^v\|^2\le\frac{\gamma_i^2}{2\mu}\|\varDelta_i\|_*^2+V(w_i^v,w_{i+1}^v).
\end{aligned}$$

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

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

Summing the above inequalities from $i=1$ *to* t*, we conclude* (82)*.*

We are now ready to prove Theorem 4.

Proof of Theorem 4. Firstly, applying (36) *and* (68) *to* (78) *in Lemma 4, we have*

$$\frac{1}{\Gamma_t}Q(w_{t+1}^{ag},u) \le \frac{\alpha_t}{\Gamma_t\gamma_t}\Omega_Z^2 + \sum_{i=1}^{t}\Lambda_i(u),\ \ \forall u \in Z. \qquad (83)\ \boxed{\texttt{eqnQLambda}}$$

Letting $w_1^v = w_1$, defining w_{i+1}^v as in (81) *with $\Delta^i = \Delta_H^{2i} + \Delta_G^i$ for all $i > 1$, we conclude from* (65) *and Lemma 5 (with $\theta_i = \alpha_i/\Gamma_i$) that*

$$-\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}+\Delta_G^i, w_i^v-u\rangle \le \mathcal{B}_t(u,w_{[t]}^v) + \sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i}+\Delta_G^i\|_*^2,\ \ \forall u \in Z. \quad (84)\ \text{?}\underline{\texttt{tmp3}}\text{?}$$

The above inequality together with (79) *and the Young's inequality yield*

$$\begin{aligned}\sum_{i=1}^{t}\Lambda_i(u) = & -\sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}+\Delta_G^i, w_i^v-u\rangle + \sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2+\|\Delta_H^{2i-1}\|_*^2\right)\\ & + \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\left[-\frac{(1-q)\mu}{2\gamma_i}\|r_i-w_{i+1}\|^2 - \langle\Delta_G^i, w_{i+1}-r_i\rangle\right]\\ & - \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_G^i, r_i-w_i^v\rangle - \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}, w_{i+1}-w_i^v\rangle\\ \le & \ \mathcal{B}_t(u,w_{[t]}^v) + U_t,\end{aligned} \qquad (85)\ \boxed{\texttt{eqnLambdaSimplifed}}$$

where

$$\begin{aligned}U_t := & \sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i}+\Delta_G^i\|_*^2 + \sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{2(1-q)\mu\Gamma_i}\|\Delta_G^i\|_*^2\\ & + \sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2+\|\Delta_H^{2i-1}\|_*^2\right)\\ & - \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_G^i, r_i-w_i^v\rangle - \sum_{i=1}^{t}\frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}, w_{i+1}-w_i^v\rangle.\end{aligned} \qquad (86)\ \boxed{\texttt{eqnU}}$$

Applying (68) *and* (85) *to* (83)*, we have*

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

or equivalently,

$$g(w_t^{ag}) \le \frac{2\alpha_t}{\gamma_t}\Omega_Z^2 + \Gamma_t U_t. \qquad (87)\ \boxed{\texttt{eqngU}}$$

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 (34), (35) *and* (81)*, during the i-th iteration of Algorithm 3, 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 (86) we have

$$\begin{aligned}\mathbb{E}[U_t] \le \mathbb{E}\Bigg[&\sum_{i=1}^t \frac{\alpha_i\gamma_i}{\mu\Gamma_i}\left(\|\Delta_H^{2i}\|^2 + \|\Delta_G^i\|_*^2\right) + \sum_{i=1}^t \frac{\alpha_i\gamma_i}{2(1-q)\mu\Gamma_i}\|\Delta_G^i\|_*^2 \\ &+ \sum_{i=1}^t \frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\left(\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2\right)\Bigg] \\ = &\sum_{i=1}^t \frac{\alpha_i\gamma_i}{\mu\Gamma_i}\left[4\sigma_H^2 + \left(1+\frac{1}{2(1-q)}\right)\sigma_G^2\right].\end{aligned} \tag*{(88) $\boxed{\texttt{eqnEUt}}$}$$

Taking expectation on both sides of (87), and using (88), we obtain (37).

*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 [18]). 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

$$Prob\left\{\sum_{i=1}^t \frac{\alpha_i}{\Gamma_i}\langle\Delta_G^i, r_i - w_i^v\rangle > \lambda\sigma_G\Omega_Z\sqrt{\frac{2}{\mu}\sum_{i=1}^t\left(\frac{\alpha_i}{\Gamma_i}\right)^2}\right\} \le \exp\{-\lambda^2/3\}. \tag*{(89) ?\underline{\texttt{tmpp1}}?}$$

By using a similar argument we have

$$Prob\left\{\sum_{i=1}^t \frac{\alpha_i}{\Gamma_i}\langle\Delta_H^{2i}, w_{i+1} - w_i^v\rangle > \lambda\sigma_H\Omega_Z\sqrt{\frac{2}{\mu}\sum_{i=1}^t\left(\frac{\alpha_i}{\Gamma_i}\right)^2}\right\} \le \exp\{-\lambda^2/3\}. \tag*{(90) ?\underline{\texttt{tmpp2}}?}$$

*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*

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

Therefore, by Markov's inequality we have

$$\begin{aligned}&Prob\left\{\left(1+\frac{1}{2(1-q)}\right)\sum_{i=1}^t\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\|\Delta_G^i\|_*^2 > (1+\lambda)\sigma_G^2\left(1+\frac{1}{2(1-q)}\right)\sum_{i=1}^t\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\right\} \\ &\le\ \exp\{-\lambda\}.\end{aligned} \tag*{(91) ?\underline{\texttt{tmpp3}}?}$$

Using similar arguments, we also have

$$Prob\left\{\sum_{i=1}^{t}\frac{3\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i-1}\|_*^2 > (1+\lambda)\frac{3\sigma_H^2}{2}\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\right\} \le \exp\{-\lambda\}, \quad (92)\ ?\texttt{tmpp4}?$$

$$Prob\left\{\sum_{i=1}^{t}\frac{5\alpha_i\gamma_i}{2\mu\Gamma_i}\|\Delta_H^{2i}\|_*^2 > (1+\lambda)\frac{5\sigma_H^2}{2}\sum_{i=1}^{t}\frac{\alpha_i\gamma_i}{\mu\Gamma_i}\right\} \le \exp\{-\lambda\}. \quad (93)\ \boxed{\texttt{tmpp5}}$$

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* (87)–(93) *that* (38) *holds.*

In the remaining part of this subsection, we will focus on proving Theorem 5, which describes the rate of convergence of Algorithm 3 for solving $SVI(Z;G,H,J)$ when Z is unbounded.

Proof the Theorem 5*. Let* U_t *be defined in* (86)*. Firstly, applying* (43) *and* (85) *to* (78) *in Lemma 4, we have*

$$\frac{1}{\Gamma_t}Q(w_{t+1}^{ag},u) \quad (94)\ \boxed{\texttt{eqnQEuclS}}$$

$$\le \mathcal{B}_t(u,r_{[t]}) - \frac{\alpha_t}{2\Gamma_t\gamma_t}\sum_{i=1}^{t}\left(q-c^2\right)\|r_i-w_{i+1}\|^2 + \mathcal{B}_t(u,w_{[t]}^v) + U_t,\ \forall u\in Z. \quad (95)\ \{?\}$$

In addition, applying (43) *to the definition of* $\mathcal{B}_t(\cdot,\cdot)$ *in* (65)*, we obtain*

$$\mathcal{B}_t(u,r_{[t]}) = \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1-u\|^2 - \|r_{t+1}-u\|^2) \quad (96)\ \boxed{\texttt{eqnBr}}$$

$$= \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1-w_{t+1}^{ag}\|^2 - \|r_{t+1}-w_{t+1}^{ag}\|^2 + 2\langle r_1-r_{t+1}, w_{t+1}^{ag}-u\rangle). \quad (97)\ \boxed{\texttt{eqnBrw}}$$

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

$$\mathcal{B}_t(u,w_{[t]}^v) = \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1-u\|^2 - \|w_{t+1}^v-u\|^2) \quad (98)\ \boxed{\texttt{eqnBv}}$$

$$= \frac{\alpha_t}{2\Gamma_t\gamma_t}(\|r_1-w_{t+1}^{ag}\|^2 - \|w_{t+1}^v-w_{t+1}^{ag}\|^2 + 2\langle r_1-w_{t+1}^v, w_{t+1}^{ag}-u\rangle). \quad (99)\ \boxed{\texttt{eqnBvw}}$$

We then conclude from (94)*,* (97)*, and* (99) *that*

$$Q(w_{t+1}^{ag},u) - \langle v_{t+1}, w_{t+1}^{ag}-u\rangle \le \varepsilon_{t+1},\ \forall u\in Z, \quad (100)\ \boxed{\texttt{tmp}}$$

where

$$v_{t+1} := \frac{\alpha_t}{\gamma_t}(2r_1 - r_{t+1} - w_{t+1}^v) \quad (101)\ \boxed{\texttt{eqnvS}}$$

and

$$\varepsilon_{t+1} := \frac{\alpha_t}{2\gamma_t}\left(2\|r_1-w_{t+1}^{ag}\|^2 - \|r_{t+1}-w_{t+1}^{ag}\|^2 - \|w_{t+1}^v-w_{t+1}^{ag}\|^2 - \sum_{i=1}^{t}\left(q-c^2\right)\|r_i-w_{i+1}\|^2\right) + \Gamma_t U_t. \quad (102)\ \boxed{\texttt{eqnepsS}}$$

It is easy to see that the residual ε_{t+1} *is positive by setting* $u = w^{ag}_{t+1}$ *in* (100)*. Hence* $\tilde{g}(w^{ag}_{t+1}, v_{t+1}) \le \varepsilon_{t+1}$*. To finish the proof, it suffices to estimate the bounds for* $\mathbb{E}[\|v_{t+1}\|]$ *and* $\mathbb{E}[\varepsilon_{t+1}]$*.*

Letting $u = u^*$ *in* (94)*, we conclude from* (96) *and* (98) *that*

$$
\begin{aligned}
&2\|r_1 - u^*\|^2 - \|r_{t+1} - u^*\|^2 - \|w^v_{t+1} - u^*\|^2 - \sum_{i=1}^{t}\left(q - c^2\right)\|r_i - w_{i+1}\|^2 + \frac{2\Gamma_t\gamma_t}{\alpha_t}U_t \\
\ge &\frac{1}{\Gamma_t}Q(w^{ag}_{t+1}, u^*) \ge 0,
\end{aligned}
$$

where the last inequality follows from (72)*. Using the above inequality and the definition of* D *in* (31)*, we have*

$$
\|r_{t+1} - u^*\|^2 + \|w^v_{t+1} - u^*\|^2 + \sum_{i=1}^{t}\left(q - c^2\right)\|r_i - w_{i+1}\|^2 \le 2D^2 + \frac{2\Gamma_t\gamma_t}{\alpha_t}U_t. \tag{103}
$$

eqnrwvBound

In addition, applying (43) *and the definition of* C_t *in* (46) *to* (88)*, we have*

$$
\mathbb{E}[U_t] \le \sum_{i=1}^{t}\frac{\alpha_t\gamma_i^2}{\Gamma_t\gamma_t}\left[4\sigma_H^2 + \left(1 + \frac{1}{2(1-q)}\right)\sigma_G^2\right] = \frac{\alpha_t}{\Gamma_t\gamma_t}C_t^2. \tag{104}
$$

eqnEUC

Combining (103) *and* (104)*, we have*

$$
\mathbb{E}[\|r_{t+1} - u^*\|^2] + \mathbb{E}[\|w^v_{t+1} - u^*\|^2] + \sum_{i=1}^{t}\left(q - c^2\right)\mathbb{E}[\|r_i - w_{i+1}\|^2] \le 2D^2 + 2C_t^2. \tag{105}
$$

eqnDC

We are now ready to prove (44)*. Observe from the definition of* v_{t+1} *in* (101) *and the definition of* D *in* (31) *that* $\|v_{t+1}\| \le \alpha_t(2D + \|w^v_{t+1} - u^*\| + \|r_{t+1} - u^*\|)/\gamma_t$*, using the previous inequality, Jensen's inequality, and* (105)*, we obtain*

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

Our remaining goal is to prove (45)*. By applying Proposition* 2 *to* (34) *and* (35) *(with* $r = r_t, w = w_{t+1}, y = r_{t+1}, \vartheta = \gamma_t\mathcal{H}(r_t, \zeta_{2t-1}) + \gamma_t\mathcal{G}(w^{md}_t, \xi_t), \eta = \gamma_t\mathcal{H}(w_{t+1}, \zeta_{2t}) + \gamma_t\mathcal{G}(w^{md}_t, \xi_t), J = \gamma_t J,\ L = 3M^2\gamma_t^2$ *and* $M^2 = 3\gamma_t^2(\|\Delta_H^{2t}\|_*^2 + \|\Delta_H^{2t-1}\|_*^2)$*), and using* (55) *and* (80)*, 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 (43). Now using (75), (76), (102), the above inequality, and applying Jensen's inequality, we have

$$
\begin{aligned}
\varepsilon_{t+1} - \Gamma_t U_t &\le \frac{\alpha_t}{\gamma_t}\|r_1 - w_{t+1}^{ag}\|^2 \\
&= \frac{\alpha_t}{\gamma_t}\left\| r_1 - u^* + \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i}(u^* - r_{i+1}) + \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i}(r_{i+1} - w_{i+1}) \right\| \\
&\le \frac{3\alpha_t}{\gamma_t}\left[D^2 + \Gamma_t \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i}\left(\|r_{i+1} - u^*\|^2 + \|w_{i+1} - r_{i+1}\|^2 \right) \right] \qquad (106)\ \boxed{\texttt{tmp2}} \\
&\le \frac{3\alpha_t}{\gamma_t}\Bigg[D^2 + \Gamma_t \sum_{i=1}^{t} \frac{\alpha_i}{\Gamma_i}\Big(\|r_{i+1} - u^*\|^2 + c^2\|w_{i+1} - r_i\|^2 \\
&\qquad + 3\gamma_i^2 (\|\Delta_H^{2i}\|_*^2 + \|\Delta_H^{2i-1}\|_*^2)\| \Big) \Bigg].
\end{aligned}
$$

Noting that by (46) and (103),

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

and that by (43),

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

we conclude from (104), (106) and Assumption ***A1.*** *that*

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

Finally, observing from (46) and (76) 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 (45) from the above inequality.

5 Numerical experiments

⟨secNumerical⟩ In this section, we present some preliminary experimental results on solving deterministic and stochastic variational inequality problems using the AMP algorithm. The comparisons with the extragradient method [15], the mirror-prox method in [23] and the stochastic mirror-prox method in [14] are provided for better examination of the performance of the AMP algorithm.

5.1 Variational inequality on the Lorentz cone

⟨secAVI⟩ In this section, we compare the performance of AMP with the extragradient/mirror-prox method [15,23], whose iteration complexity has been studied in [23,20]. In particular, we consider an affine variational inequality problem with unbounded feasible set on solving $u^* \in Z$ such that

$$\langle Au + b, u^* - u\rangle \le 0,\ \forall u \in Z, \tag{107}$$ eqnAVI

where $A \in \mathbb{R}^{(n+1)\times(n+1)}$ is a linear monotone operator, and Z is the Lorentz cone:

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

To solve (107), 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 VI problem (107) can be viewed as an instance of (1) with

$$F(u) = Au + b,\ G(u) = \frac{1}{4}\langle (A + A^T)u, u\rangle + \langle b, u\rangle,\ H(u) = \frac{1}{2}(A - A^T)u, \tag{108}$$ eqnAVISetting

and $J(u) = 0$.

A few remarks are in place for the above decomposition. Firstly, for any continuous linear monotone operators on Banach spaces, the decomposition to the sum of a symmetric monotone operator (and hence the subdifferential of a convex function) and a skew operator exists and is unique (see, e.g., Proposition 2.14 in [3]). Therefore, it is natural to use the decomposition (108) to solve (107). Secondly, as discussed after Algorithm 1, the AMP algorithm can be viewed as a hybrid algorithm of the mirror-prox method and the accelerated gradient method. Indeed, the AMP algorithm is equivalent to the mirror-prox method and a version of Nesterov's accelerated method when A is skew-symmetric and symmetric, respectively. Finally, since the Lipschitz constants of $F(\cdot)$, $\nabla G(\cdot)$ and $H(\cdot)$ are $\|A\|$, $\|A + A^T\|/2$ and $\|A - A^T\|/2$ respectively, the iteration complexity of the mirror-prox method for computing an approximate solution of (107) is $\mathcal{O}(\|A\|/\varepsilon)$, and the iteration complexity of the AMP method is $\mathcal{O}(\|A + A^T\|/\sqrt{\varepsilon} + \|A - A^T\|/\varepsilon)$. Specially, if A is "almost symmetric", i.e., $\|A\|$ is much greater than the norm of its skew-symmetric part $\|A - A^T\|/2$ (e.g., $\|A\| \ge \|A - A^T\|/\sqrt{\varepsilon}$), then the iteration complexity of the AMP method for computing an approximate solution of (107) is better than that of the mirror-prox method (e.g., in the order of $\mathcal{O}(1/\sqrt{\varepsilon})$).

In this experiment, we generate the linear monotone operator A randomly by $A = B^TB + (C - C^T)$, where $B \in \mathbb{R}^{\lceil (n+1)/2\rceil \times (n+1)}$ (so that A is monotone by 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. The entries of the vector b are also randomly distributed between 0 and 1. By setting $V(z,u) = \|z-u\|^2/2$, the prox-mapping $P_z^J(u)$ in (11) becomes the projection of $z-\eta$ to the Lorentz cone Z, which can be calculated efficiently. For the AMP algorithm, we use the parameter settings in Corollary 1 with $L = \|A+A^T\|/2$ and $M = \|A-A^T\|/2$, and for the extragradient method we choose the stepsizes according to (3.2) of [23] in which $L = \|A\|$. Noting the fact that AMP computes relatively more matrix-vector multiplications due to the aforementioned decomposition, we set the total number of iterations of the extragradient method to be twice of that of the AMP method. The performance of the AMP and extragradient algorithms are compared in terms of the gap function (28), which is computed using MOSEK [22]. In particular, for any approximate solution w and perturbation vector v, we compute the value of $\tilde{g}(w,v)$ in (28) and the norm of the perturbation vector $\|v\|$.[1] It should be noted that in this experiment both the AMP and extragradient methods are implemented without using backtracking procedures in order to have a fair comparison. The comparison between the AMP algorithm and the extragradient algorithm is described in Table 1.

⟨tabAVI⟩ **Table 1** The comparison of the AMP algorithm and the extragradient (denoted by EG) algorithm in solving the affine variational inequality problem (107). 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).

Problem	N	AMP, N iterations			EG, $2N$ iterations		
		$\tilde{g}(w,v)$	$\|v\|$	CPU	$\tilde{g}(w,v)$	$\|v\|$	CPU
$n = 999$, $L = 2872.3, M = 25.9$	1000	7.91e-3	1.03e-1	1.3	2.69e-2	1.12e0	1.2
	2000	3.63e-3	4.87e-2	2.5	2.09e-2	6.05e-1	2.3
4999 $L = 14472.8, M = 57.6$	1000	1.84e-1	7.10e-1	34.2	8.26e-2	6.30e0	46.0
	2000	7.91e-2	3.20e-1	69.5	1.39e-1	4.89e0	91.7
9999 $L = 29056.0, M = 81.4$	1000	1.86e-1	8.88e-1	142.1	7.16e-2	8.46e0	192.9
	2000	7.60e-2	3.85e-1	286.0	1.13e-1	6.69e0	379.3

Two remarks on the performance of the AMP and extragradient methods 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 AMP algorithm outperforms the extragradient method for solving (107). This is consistent with our theoretical observation that the AMP algorithm has a better iteration complexity bound than that of the extragradient method for solving problem (107). Especially, it can be easily seen the performance of the AMP method on the perturbation vector $\|v\|$ is significantly better than that of the extragradient method.

5.2 Multi-player nonlinear game

⟨secnGame⟩ The goal of this section is to compare the AMP algorithm with backtracking and the mirror-prox algorithm with adaptive stepsizes in [23]. More specifically, we calculate the Nash equilibrium of a game among k players, in which the goal of each player is to minimize his/her quadratic loss function. To model the game, the strategies of

[1] See the proof of Theorem 3 for the definition of the perturbation term in the AMP algorithm, and Theorem 5.2 in [20] for the definition of the perturbation term in the extragradient algorithm.

the players are denoted by $x_1, \dots, x_k$, which represent the portfolio investment of the players and are described as a point on the standard simplex, i.e.,

$$x_i \in \Delta^n := \left\{ x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^{(i)} = 1 \right\}, \ \forall i = 1, \dots, k.$$

Then, the loss function of the i-th player is modeled by

$$\phi_i(x_1, \dots, x_k) = \frac{1}{2}\langle A_{i,i}x_i, x_i\rangle + \sum_{j \neq i}\langle x_i, A_{i,j}x_j\rangle.$$

There are two types of losses in the above function, i.e., the first term describes the impact of strategy x_i on the i-player him/herself, and the second term describes the outcome of the i-th player's strategy x_i when interacting with the strategies of other players. We assume that $A_{i,j} = -A_{j,i}^T$ so that the pairwise interactions between any two players i and j results in a zero-sum outcome, and that $A_{i,i}$ is positive semidefinite for all i so that our multi-player game is convex (For the detailed introduction of solving the multi-player game using variational inequalities, see, e.g., [24, 14]).

By [24], the Nash equilibrium of the above game is exactly the weak solution of the VI problem (1), where $Z = \Delta^n \times \ldots \times \Delta^n$ is the k-product space of all possible collections of strategies, and for all $z = (x_1^T, \dots, x_k^T)^T$,

$$F(z) = Az \text{ with } A := \begin{pmatrix} A_{1,1} & \cdots & A_{1,k} \\ \vdots & \ddots & \vdots \\ A_{k,1} & \cdots & A_{k,k} \end{pmatrix}. \qquad (109)\ \boxed{\texttt{eqnFBlock}}$$

Similarly to Section 5.1, we consider a decomposition of $F = \nabla G + H + J$ with

$$G(z) := \frac{1}{4}\langle (A + A^T)z, z\rangle, \ H(z) := \frac{1}{2}(A - A^T)z, \text{ and } J(z) \equiv 0.$$

To compute a solution of (1), we consider the following entropy setting for the prox-function used in the AMP algorithm: for all $z = (x_1^T, \dots, x_k^T)^T \in Z$, $u = (y_1^T, \dots, y_k^T)^T \in Z$ and $\xi = (\eta_1^T, \dots, \eta_k^T)^T \in \mathcal{E}$, we define

$$\begin{aligned} &\|z\| := \sqrt{\sum_{i=1}^k \|x_i\|_1^2}, \ \|\xi\|_* := \sqrt{\sum_{i=1}^k \|\eta_i\|_\infty^2}, \text{ and} \\ &V(z, u) := \sum_{i=1}^k \sum_{j=1}^n (y_i^{(j)} + \nu/n) \ln \frac{y_i^{(j)} + \nu/n}{x_i^{(j)} + \nu/n}. \end{aligned} \qquad (110)\ \boxed{\texttt{eqnnGameSetting}}$$

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 (11) can be efficiently solved within machine accuracy, and the strong convexity parameter of the prox-function $V(z, u)$ is $\mu = 1 + \nu$ (See [4] for details on the entropy prox-functions). Moreover, it should be noted that under the above definition of $G(z)$ and $H(z)$, for any approximate solution $u = (y_1^T, \dots, y_k^T)^T$, the gap function $g(u)$ in (18) becomes

$$g(u) = \sum_{i=1}^k \left[\phi_i(u) - \min_{x_i \in \Delta^n} \phi_i(y_1, \dots, y_{i-1}, x_i, y_{i+1}, \dots, y_k) \right],$$

⟨tabnGame⟩ **Table 2** The comparison of the AMP algorithm and the mirror-prox (denoted by MP) algorithm in computing the equilibirum of multi-player games. k is the number of players in the game, and n is the number of portfolio investments that describes the strategy of each player.

Problem dimension	AMP after 500 iterations		MP after 1000 iterations	
	$g(u)$	CPU	$g(u)$	CPU
1000 ($k = 5$, $n = 200$)	3.54e-4	1.0	3.02e-3	1.6
4000 ($k = 20$, $n = 200$)	1.63e-3	15.2	2.17e-2	21.5
10000 ($k = 50$, $n = 200$)	1.09e-2	96.2	9.61e-2	117.3

which is exactly the natural error estimate of Nash equilibria.

The matrix A in (109) is generated randomly by the following means: Firstly, for all $i < j$, the entries of $A_{i,j}$ are independently generated from the uniform $[0, 1]$ distribution, and then $A_{j,i}$ is set to $-A_{i,j}^T$. Secondly, for all i, $A_{i,i} = B_i^T B_i$ where $B_i \in \mathbb{R}^{\lceil n/2 \rceil \times n}$ and each entry of $B_{i,i}$ are independently generated from the uniform $[0, 1]$ distribution. Finally, for simplicity of this experiment, we rescale the matrices, so that for any $i \neq j$ the entry of $A_{i,j}$ with the maximum absolute value is 1, and for any i the entry of $A_{i,i}$ with the maximum absolute value is 10.[2] We compare the performance between the AMP algorithm with backtracking in Algorithm 2 and the mirror-prox algorithm with adaptive stepsizes in [23]. For any approximate solution u, we evaluate its accuracy by estimating the gap function $g(u)$, which is computed using MOSEK [22].

The comparison between the computational performance of AMP and MP is displayed in Table 2. We can see that AMP outperforms MP for solving the aforementioned multi-player game. This is consistent with our theoretical observations on the iteration complexities of AMP and MP.

5.3 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 saddle point problem

$$\min_{x \in \Delta^n} \max_{y \in \Delta^n} \frac{1}{2}\langle Px, x\rangle + \langle Kx, y\rangle - \frac{1}{2}\langle Qy, y\rangle, \tag{111}$$ eqn2Game

where P and Q are positive semidefinite matrices, and Δ^n is a standard simplex. Problem (111) is a special case of the problem in Section 5.2 with only two players. For simplicity, we only consider the case when $\max_{i,j} |P^{(i,j)}| = \max_{i,j} |Q^{(i,j)}|$ (see the footnote 2). Letting $Z := \Delta^n \times \Delta^n$, the above problem is equivalent to the VI problem (1) with

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

⟨noteScaling⟩ [2] When the maximum absolute values of $A_{i,i}$'s are different, it is recommended to introduce weights ω_i's and set $\|z\| := \sqrt{\sum_{i=1}^k \omega_i \|x_i\|_1^2}$ and $\|\xi\| := \sqrt{\sum_{i=1}^k \omega_i^{-1} \|\eta_i\|_1^2}$, in which ω_i's depend on the blocks $A_{i,j}$'s. See "mixed setups" in Section 5 of [23] for the detailed derivations.

where $u := (x, y) \in Z$. If P, Q and K are dense and n is large, the matrix-vector multiplication of Px, Qy, $K^T y$ and Kx may be very expensive. In order to reduce the arithmetic cost of computing these matrix-vector multiplications, Nemirovski et al. [25] developed randomized algorithms for solving this type of VI problems by replacing the calculations of matrix-vector multiplications with calls to a stochastic oracle. Using similar ideas to [25], we assume that for each input $(x_i, y_i) \in Z$, the $\mathcal{SO}$ outputs the *stochastic gradients* $(\hat{\mathcal{G}}_x(x_i), \hat{\mathcal{G}}_y(y_i), \hat{\mathcal{K}}_x(x_i), \hat{\mathcal{K}}_y(y_i)) \equiv (\mathcal{G}_x(x_i, \xi_i), \mathcal{G}_y(y_i, \xi_i), \mathcal{K}_x(x_i, \xi_i), \mathcal{K}_y(y_i, \xi_i))$ such that for all $j, k, l, m = 1, \ldots n$.

$$\begin{aligned}
&\text{Prob}(\hat{\mathcal{G}}_x(x_i) = P_j) = x_i^{(j)}, \ \text{Prob}(\hat{\mathcal{G}}_y(y_i) = Q_k) = y_i^{(k)}, \\
&\text{Prob}(\hat{\mathcal{K}}_x(x_i) = K_l) = x_i^{(l)}, \ \text{and} \ \text{Prob}(-\hat{\mathcal{K}}_y(y_i) = -K^m) = y_i^{(m)},
\end{aligned}$$

Here, we denote by K_l and $(K^m)^T$ the l-th column and m-th row of K, respectively. In other words, each call to the $\mathcal{SO}$ outputs the random samples of the columns of P and Q and columns and rows of K whose distributions depend on the input (x_i, y_i). It can be checked that $\mathbb{E}[\hat{\mathcal{G}}_x(x_i)] = Px_i$, $\mathbb{E}[\hat{\mathcal{G}}_y(y_i)] = Qy_i$, $\mathbb{E}[-\hat{\mathcal{K}}_x(x_i)] = -Kx_i$ and $\mathbb{E}[\hat{\mathcal{K}}_y(y_i)] = K^T y_i$. Since problem (111) is a special case of the multi-player game in Section 5.2, we still apply the entropy prox-function setting in (110). It is easy to check that

$$\begin{aligned}
&L \le \max\{\max_{k,j} |P^{(k,j)}|, \max_{k,j} |Q^{(k,j)}|\}, \ M \le \max_{k,j} |K^{(k,j)}|, \ \Omega_Z^2 = 2(1 + \frac{\nu}{n}) \ln(\frac{n}{\nu} + 1), \\
&\mathbb{E}\left[\left\| \begin{pmatrix} \hat{\mathcal{G}}_x(x_i) - Px \\ \hat{\mathcal{G}}_y(y_i) - Qy \end{pmatrix} \right\|_*^2 \right] \le 4 \left(\max_{k,j} |P^{(k,j)}|^2 + \max_{k,j} |Q^{(k,j)}|^2 \right), \ \text{and} \\
&\mathbb{E}\left[\left\| \begin{pmatrix} -\hat{\mathcal{K}}_x(x_i) + Kx \\ \hat{\mathcal{K}}_y(y_i) - K^T y \end{pmatrix} \right\|_*^2 \right] \le 8 \max_{k,j} |K^{(k,j)}|^2.
\end{aligned}$$

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 $\max_{k,j} |P^{(k,j)}| = \max_{k,j} |Q^{(k,j)}|$. For the SAMP algorithm, we use the scheme in Algorithm 3 with the parameters described in (110) and Corollary 3. As a comparison, we also implement the stochastic mirror-prox (SMP) method described in (3.6) and (3.7) in [14]. Noticing that both the SAMP and SMP algorithms are robust with respect to the above estimates of σ and Ω_Z (see the discussion after Corollary 3, and also the proof of Corollary 4.2 in [14]), we run both algorithms twice with and without fine-tuning for each problem instances. In the first run without fine-tuning, we set $\beta = \sigma/\Omega_Z$ in the SAMP algorithm and use the aforementioned stepsize constants for the SMP algorithm. In the second run, we fine-tune the value of β in the implementation of the SAMP algorithm. Specifically, for each $\varrho = 2^{-9}, 2^{-8}, \ldots, 2^8, 2^9$, we run 50 iterations of the SAMP algorithm with $\beta = \varrho\sigma/\Omega_Z$, and choose the best value β for the SAMP algorithm implementation by

⟨tab2Game⟩ **Table 3** The comparison of the SAMP and SMP algorithms in computing the equilibirum of two-player games, in terms of the expectation and standard deviation of the gap function value $g(u)$ for any approximate solution u. The CPU time in the table is the average time of 100 runs.

Problem	Paramters		SAMP			SMP		
dimension n	β	N	$\mathbb{E}[g(u)]$	std.	CPU	$\mathbb{E}[g(u)]$	std.	CPU
1000 $(L = 184.7,$ $M = 1.0,$ $\sigma = 89.4,$ $g(u_0) = 1.37e1)$	$\beta = \frac{\sigma}{\Omega_Z}$	1000	3.01e-1	8.23e-2	0.4	1.31e1	6.25e-4	0.4
		2000	2.15e-1	4.96e-2	0.7	1.29e1	6.16e-4	0.8
		5000	1.38e-1	3.37e-2	1.7	1.24e1	6.84e-4	2.1
	tuned	1000	7.32e-1	1.49e-1	0.4	5.89e0	1.67e-2	0.4
		2000	5.72e-1	1.04e-2	0.8	4.31e0	1.29e-2	0.8
		5000	3.99e-1	7.21e-2	2.4	2.73e0	8.00e-3	2.1
2000 $(L = 366.4,$ $M = 1.0,$ $\sigma = 54.2,$ $g(u_0) = 2.07e1)$	$\beta = \frac{\sigma}{\Omega_Z}$	1000	6.98e-1	1.84e-1	0.5	2.01e1	4.87e-4	0.6
		2000	4.89e-1	1.14e-1	0.9	1.99e1	4.24e-4	1.1
		5000	3.09e-1	8.18e-2	2.3	1.93e1	5.00e-4	2.7
	tuned	1000	1.03e0	2.12e-1	0.5	5.40e0	3.83e-2	0.6
		2000	7.46e-1	1.76e-1	0.9	2.65e0	1.92e-2	1.1
		5000	4.92e-1	9.32e-2	2.3	1.02e0	5.83e-3	2.7
5000 $(L = 893.5,$ $M = 1.0,$ $\sigma = 84.6,$ $g(u_0) = 3.56e1)$	$\beta = \frac{\sigma}{\Omega_Z}$	1000	1.70e0	3.06e-3	0.9	3.51e1	3.19e-4	1.1
		2000	1.22e0	2.25e-1	1.8	3.48e1	3.14e-4	2.1
		5000	7.80e-1	1.27e-1	4.4	3.43e1	3.34e-4	5.2
	tuned	1000	3.89e0	5.98e-1	1.0	1.49e1	4.24e-2	1.1
		2000	2.83e0	3.93e-1	1.9	7.63e0	2.92e-2	2.1
		5000	2.96e0	2.99e-1	4.6	4.00e0	1.47e-2	5.2

comparing the gap function values (18). The same fine-tuning strategy is also applied to the SMP algorithm as it is robust with respect to the value of $M/\sqrt{\Theta}$ in (4.3) in [14]. The performance of the SAMP and SMP algorithms are compared in terms of the mean and standard deviation of the gap function values (18) (computed by MOSEK [22]) in 100 runs.

The comparison between the SAMP and SMP algorithms in terms of the performance on computing approximate solutions of (111) 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.

6 Conclusion

⟨secConclusion⟩ We present in this paper a novel accelerated mirror-prox (AMP) method for solving a class of deterministic and stochastic variational inequality (VI) problems. The basic idea of this algorithm is to incorporate a multi-step acceleration scheme into the mirror-prox method in [23, 14]. For both the deterministic and stochastic VI, the AMP 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. Moreover, the iteration cost of the AMP is comparable to, or even less than that of the mirror-prox method in that it saves one compuation of $\nabla G(\cdot)$. To the best of our knowledge, this is the first algorithm with the optimal iteration complexity bounds for solving the deterministic and stochastic VIs of type (2). Furthermore, we show that the developed AMP 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 AMP depends on the distance from the initial point to the set of strong solutions. Specially, in the unbounded case of the deterministic VI, the AMP scheme achieves the iteration complexity without requiring any knowledge on the distance from the initial point to the set of strong solutions. Our preliminary numerical results show that the proposed AMP algorithm is promising to solve large-scale variational inequality problems.

References

ender2005interior 1. A. Auslender and M. Teboulle, *Interior projection-like methods for monotone variational inequalities*, Mathematical programming, 104 (2005), pp. 39–68.

ender2006interior 2. ———, *Interior gradient and proximal methods for convex and conic optimization*, SIAM Journal on Optimization, 16 (2006), pp. 697–725.

hke1995continuous 3. H. H. Bauschke and J. M. Borwein, *Continuous linear monotone operators on banach spaces*, (1995).

ben-tal2005non 4. A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Mathematical Programming, 102 (2005), pp. 407–456.

ekas1999nonlinear 5. D. P. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.

man1967relaxation 6. L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR computational mathematics and mathematical physics, 7 (1967), pp. 200–217.

ik1997enlargement 7. R. S. Burachik, A. N. Iusem, and B. F. Svaiter, *Enlargement of monotone operators with applications to variational inequalities*, Set-Valued Analysis, 5 (1997), pp. 159–180.

chen1999homotopy 8. X. Chen and Y. Ye, *On homotopy-smoothing methods for box-constrained variational inequalities*, SIAM Journal on Control and Optimization, 37 (1999), pp. 589–616.

chen2013optimal 9. Y. Chen, G. Lan, and Y. Ouyang, *Optimal primal-dual methods for a class of saddle point problems*, http://www.optimization-online.org/DB_HTML/2013/04/3850.html, (2013).

g2012convergence 10. C. D. Dang and G. Lan, *On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators*, arXiv preprint arXiv:1311.2776, (2013).

chinei2003finite 11. F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer, 2003.

adimi2012optimal 12. S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework*, SIAM Journal on Optimization, 22 (2012), pp. 1469–1492.

adimi2013optimal 13. ———, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 2061–2089.

itsky2011solving 14. A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stochastic Systems, 1 (2011), pp. 17–58.

976extragradient 15. G. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.

983extrapolation 16. ———, *Extrapolation gradient methods and relation to modified Lagrangians*, Ekonomika i Matematicheskie Metody, 19 (1983), pp. 694–703. in Russian; English translation in Matekon.

lan2012optimal 17. G. Lan, *An optimal method for stochastic composite optimization*, Mathematical Programming, 133 (1) (2012), pp. 365–397.

an2012validation 18. G. Lan, A. Nemirovski, and A. Shapiro, *Validation analysis of mirror descent stochastic approximation method*, Mathematical programming, (2012).

70regularisation 19. B. Martinet, *Regularisation d'inéquations variationelles par approximations successives*, Revue Française d'Automatique, Informatique et Recherche Opérationnelle, 4 (1970), pp. 154–159.

ro2010complexity 20. R. D. Monteiro and B. F. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM Journal on Optimization, 20 (2010), pp. 2755–2787.

ro2011complexity 21. ———, *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 Journal on Optimization, 21 (2011), pp. 1688–1720.

mosek2012mosek 22. A. MOSEK, *The mosek optimization toolbox for matlab manual, version 6.0 (revision 135)*, MOSEK ApS, Denmark, (2012).

emirovski2005prox 23. A. Nemirovski, *Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM Journal on Optimization, 15 (2004), pp. 229–251.

vski2010accuracy 24. A. Nemirovski, *Accuracy certificates for computational problems with convex structure*, . . . of Operations Research, 35 (2010), pp. 52–78.

irovski2009robust 25. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.

ovski1983problem 26. A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley, XV, 1983.

ki1992information 27. A. S. Nemirovski, *Information-based complexity of linear operator equations*, Journal of Complexity, 8 (1992), pp. 153–175.

sterov1983method 28. Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence* $O(1/k^2)$, Doklady AN SSSR, 269 (1983), pp. 543–547. translated as Soviet Math. Docl.

esterov2005smooth 29. ———, *Smooth minimization of non-smooth functions*, Mathematical Programming, (2005), pp. 1–26.

nesterov2007dual 30. ———, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Mathematical Programming, 109 (2007), pp. 319–344.

esterov2009primal 31. ———, *Primal-dual subgradient methods for convex problems*, Mathematical programming, 120 (2009), pp. 221–259.

v1999homogeneous 32. Y. Nesterov and J. P. Vial, *Homogeneous analytic center cutting plane methods for convex problems and variational inequalities*, SIAM Journal on Optimization, 9 (1999), pp. 707–728.

ellar1976monotone 33. R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

bony1970methodes 34. M. Sibony, *Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone*, Calcolo, 7 (1970), pp. 65–183.

lodov1999hybrida 35. M. V. Solodov and B. F. Svaiter, *A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Analysis, 7 (1999), pp. 323–345.

olodov1999hybrid 36. ———, *A hybrid projection-proximal point algorithm*, Journal of convex analysis, 6 (1999), pp. 59–70.

solodov1999new 37. ———, *A new projection method for variational inequality problems*, SIAM Journal on Control and Optimization, 37 (1999), pp. 765–776.

lodov2000inexact 38. ———, *An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions*, Mathematics of Operations Research, 25 (2000), pp. 214–230.

sun1995new 39. D. Sun, *A new step-size skill for solving a class of nonlinear projection equations*, Journal of Computational Mathematics, 13 (1995), pp. 357–368.

seng2000modified 40. P. Tseng, *A modified forward-backward splitting method for maximal monotone mappings*, SIAM Journal on Control and Optimization, 38 (2000), pp. 431–446.

g2008accelerated 41. P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, submitted to SIAM Journal on Optimization, (2008).