AN EXTRA GRADIENT ANDERSON-ACCELERATED ALGORITHM FOR PSEUDOMONOTONE VARIATIONAL INEQUALITIES AND ITS EXTENSION TO STOCHASTIC PROBLEMS

XIN QU

PhD

The Hong Kong Polytechnic University

This programme is jointly offered by The Hong Kong Polytechnic University and Harbin Institute of Technology

2026

The Hong Kong Polytechnic University
Department of Applied Mathematics

Harbin Institute of Technology
Department of Mathematics

An Extra Gradient Anderson-Accelerated Algorithm for Pseudomonotone Variational Inequalities and Its Extension to Stochastic Problems

Xin Qu

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

October, 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Xin Qu (Name of student)

Dedicated to my parents and grandparents.

Abstract

Nonmonotone variational inequalities represent a challenging class of problems that arise in various applications, including economics, engineering, and network analysis, where traditional monotonicity assumptions do not hold. The study of algorithms for nonmonotone variational inequalities has consistently been a central topic in optimization. This thesis aims to design efficient algorithms for solving pseudomonotone variational inequality problems, which are a special class of nonmonotone variational inequalities. The content is mainly divided into the following two aspects.

In the first part of the thesis, we propose an extra gradient Anderson-accelerated algorithm for solving pseudomonotone variational inequalities, which uses the extra gradient scheme with line search to guarantee the global convergence and Anderson acceleration to achieve fast convergent rate. We prove that the sequence generated by the proposed algorithm from any initial point converges to a solution of the pseudomonotone variational inequality problem without assuming the Lipschitz continuity and contractive condition, which are used for convergence analysis of the extra gradient method and Anderson-accelerated method, respectively in existing literatures. Subsequently, under the condition that the operator is locally Lipschitz continuous, we provide the convergence rate concerning the residual function. Additionally, we propose an improved version of the above algorithm, aiming to enhance convergence efficiency by reducing the number of projections onto the feasible set. Numerical experiments, particular emphasis on Harker-Pang problems, fractional

programming problems, nonlinear complementarity problems, PDE problems with free boundary and linear complementarity problems, are conducted to validate the effectiveness and good performance of the proposed algorithm comparing with the existing extra gradient method and Anderson-accelerated method.

In the second part of the thesis, we propose a stochastic extra gradient Anderson-accelerated algorithm for solving pseudomonotone stochastic variational inequalities. This algorithm is developed based on the extra gradient method and Anderson-accelerated method, utilizing a stochastic approximation approach. By incorporating a line search technique to against the unknown Lipschitz constants, the proposed stochastic algorithm is designed. We prove that the sequence generated by the proposed stochastic algorithm converges almost surely to a solution of the pseudomonotone stochastic variational inequality problem. Additionally, we establish the sublinear convergence rate of the proposed stochastic algorithm in terms of the mean residual function, along with its optimal oracle complexity. Finally, we validate the effectiveness of the proposed stochastic algorithm by some experiments.

This thesis contains research results of the following paper which has been published online during the period of my Ph.D study at the Department of Applied Mathematics, The Hong Kong Polytechnic University:

- X. QU, W. BIAN and X. CHEN, *An extra gradient Anderson-accelerated algorithm for pseudomonotone variational inequalities*, to appear in Mathematics of Computation.

Acknowledgements

The years spent pursuing my doctoral degree have been an incredibly important and memorable period in my life. I would like to express my heartfelt gratitude to those who have provided me with support and strength throughout my academic journey.

First and foremost, I would like to express my gratitude to my supervisor, Prof. Xiaojun Chen. Throughout the research process, Prof. Xiaojun Chen consistently provided me with valuable advice and guidance, encouraging me to explore boldly and innovate courageously. She helped me continuously refine my research methods and broaden my perspectives. Her rigorous academic attitude and pursuit of excellence in research have deeply inspired me. Her attention to detail and high standards for research quality motivate me to keep moving forward on my academic journey.

I would like to express my gratitude to my supervisor at Harbin Institute of Technology, Prof. Wei Bian. Her dedicated guidance and extensive knowledge have been a constant source of inspiration for me. Prof. Wei Bian has been a guiding light on my academic journey. Her expertise, passion for academia, and rigorous attitude have profoundly influenced me, providing the motivation I need to progress.

I would like to extend my gratitude to all members of Prof. Chen's research group: Prof. Chao Zhang, Prof. Xin Liu, Prof. Hailin Sun, Prof. Yanfang Zhang, Prof. Zaikun Zhang, Prof. Xiao Wang, Prof. Yang Zhou, Prof. Bo Wen, Prof. Jie Jiang, Prof. Lei Yang, Prof. Hong Wang, Prof. Jianfeng Luo, Dr. Fang He, Dr. Shisen Liu, Dr. Xiaozhou Wang, Dr. Xiaoxia Liu, Dr. Fei Fang, Dr. You Zhao, Dr.

Yong Zhao, Dr. Zhihua Zhao, Dr. Lin Chen, Dr. Kaixin Gao, Dr. Ming Huang, Dr. Guan Wang, Dr. Lei Wang, Dr. Zicheng Qiu, Dr. Shijie Yu, Dr. Yifan He, Dr. Yue Wang, Mr. Xiao Zha, Mr. Zhouxing Luo, Mr. Wentao Ma, Miss Yixuan Zhang, Miss Lingzi Jin, Miss Xue Li and Miss Yiyang Li. In particular, I am deeply grateful to Dr. Xingbang Cui, Dr. Chao Li, Dr. Fan Wu and Dr. Yi Zhang for thoroughly reviewing my thesis and offering insightful feedback. Their guidance and suggestions have significantly enhanced the quality of this work. I would like to express my sincere gratitude to all the friends I have met at PolyU. Their companionship and support have ensured that I was never alone on my academic journey.

I would like to express my deepest gratitude to my family for their unwavering support throughout my academic journey. Their unconditional love has been a source of immense strength for me. I am especially thankful to my parents, whose encouragement and belief in me have helped me face the challenges along the way. I am truly blessed to have such a supportive family by my side.

Contents

Certificate of Originality **iii**

Abstract **vii**

Acknowledgements **xi**

List of Figures **xv**

List of Tables **xvii**

List of Notation **xix**

1 Introduction **1**

1.1 Background 1

1.2 Literature reviews 2

1.2.1 The projection methods for solving variational inequalities . . 3

1.2.2 Stochastic approximation methods for solving stochastic variational inequalities 6

1.2.3 Anderson acceleration for fixed point problems 9

1.3 Contribution of the thesis 11

1.4 Organization of the thesis 13

2 Basic Notation and Preliminaries **15**

2.1 Basic notation 15

2.2 Preliminaries 16

3 An Extra Gradient (EG) Anderson-Accelerated Algorithm for Pseudomonotone Variational Inequalities 19

3.1 EG-Anderson(1) algorithm and its convergence analysis 20

3.1.1 Framework of the EG-Anderson(1) algorithm 20

3.1.2 Convergence analysis 22

3.2 Subgradient EG-Anderson(1) algorithm and its convergence analysis . 34

3.2.1 Framework of the subgradient EG-Anderson(1) algorithm . . . 35

3.2.2 Convergence analysis 35

3.3 Numerical experiments 41

4 A Stochastic Accelerated Algorithm for Pseudomonotone Variational Inequalities via Anderson Acceleration 55

4.1 Stochastic EG-Anderson(1) algorithm and its convergence analysis . . 56

4.1.1 Framework of the stochastic EG-Anderson(1) algorithm 56

4.1.2 Convergence analysis 61

4.2 Numerical experiments 82

4.2.1 Simple examples 82

4.2.2 Portfolio management 87

5 Conclusions and Future Work 93

5.1 Conclusions 93

5.2 Future work 94

Bibliography 97

List of Figures

3.1 Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.1 43

3.2 Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.2 45

3.3 Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.3 47

3.4 Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.4 50

4.1 Comparisons of the convergence behaviours of the stochastic EG-Anderson(1) and the stochastic EG algorithms for Example 4.1 . . . 84

4.2 Comparisons of the convergence behaviours of the stochastic EG-Anderson(1) and the stochastic EG algorithms for Example 4.2 . . . 86

4.3 Convergence curves and comparison of SR and CR on the testing set of Data 2 92

List of Tables

1.1 Summary of results on algorithms with global convergence to nonmonotone VIs 7

3.1 Comparisons of the three algorithms for Example 3.1 44

3.2 Comparisons of the three algorithms for Example 3.2 46

3.3 Comparisons of the three algorithms for Example 3.3 48

3.4 Comparisons of the three algorithms for Example 3.4 with $p = 0.9$ . . 51

3.5 Comparisons of the three algorithms for Example 3.4 with $p = 0.8$ . . 52

3.6 Comparisons of the four algorithms for Example 3.5 53

4.1 Comparisons of the two algorithms for Example 4.1 85

4.2 Comparisons of the two algorithms for Example 4.2 87

4.3 Description of the six real data sets 89

4.4 Performance metrics for different data sets 92

List of Notation

$\mathbb{R}$	the set of real numbers
$\mathbb{R}^n$	the set of n-dimensional real vectors
$\mathbb{R}^{m\times n}$	the set of $m \times n$ real matrices
$\mathbb{R}^n_+$	the n-dimensional nonnegative orthant
$\mathbb{N}$	the set of positive integers
$\mathbb{N}_0$	the set of nonnegative integers
$[m]$	for a given $m \in \mathbb{N}$, $[m] := \{1, 2, \ldots, m\}$
$\lceil a \rceil$	the smallest integer greater than or equal to $a \in \mathbb{R}$
e	the base of the natural logarithm
$\mathbf{e}$	the vector of all ones in $\mathbb{R}^n$
$\mathbf{0}$	the vector of all zeros in $\mathbb{R}^n$
$x^\top$	the transpose of a vector x
$A^\top$	the transpose of a matrix A
$\|x\|$	the ℓ_2-norm of a vector x
$\|A\|$	the ℓ_2-norm of a matrix A
$\|A\|_F$	the Frobenius norm of a matrix A
$d(x, \Omega)$	the distance from a vector x to the set Ω, i.e., $d(x, \Omega) := \inf\{\|x - y\| : y \in \Omega\}$
$P_\Omega(x)$	the projection of x onto the set Ω

$\min(x, y)$	the vector with its i-th component is $\min(x_i, y_i)$
$\max(x, y)$	the vector with its i-th component is $\max(x_i, y_i)$
$\nabla f(x)$	the gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$
$(X, \mathcal{F}, \mathbb{P})$	a probability space, where X is a sample space, $\mathcal{F}$ is a σ-algebra, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure
ξ	a random variable defined in the probability space $(X, \mathcal{F}, \mathbb{P})$
$\sigma(\xi^1, \cdots, \xi^k)$	the σ-algebra generated by the random variables $\xi^1, \cdots, \xi^k$
$\mathbb{E}[\xi]$	the expectation of ξ
$\mathbb{E}[\xi\|\mathcal{F}]$	the conditional expectation of ξ with respect to $\mathcal{F}$
$\|\xi\|_p$	the L^p-norm of ξ for $p \geq 1$, i.e., $\|\xi\|_p := (\mathbb{E}[\|\xi\|^p])^{\frac{1}{p}}$
$\|\xi\|\mathcal{F}\|_p$	the L^p-norm of ξ conditional to $\mathcal{F}$ for $p \geq 1$, i.e., $\|\xi\|\mathcal{F}\|_p := (\mathbb{E}[\|\xi\|^p\|\mathcal{F}])^{\frac{1}{p}}$
i.i.d.	independent identically distributed
a.s.	almost surely
$\mathrm{VI}(\Omega, H)$	the variational inequality problem defined by the feasible set Ω and operator H
$\mathrm{SOL}(\Omega, H)$	the solution set of $\mathrm{VI}(\Omega, H)$
$\mathrm{SVI}(\Omega, \mathbb{E}[T(\xi, \cdot)])$	the stochastic variational inequality problem defined by the feasible set Ω and operator T
$\mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$	the solution set of $\mathrm{SVI}(\Omega, \mathbb{E}[T(\xi, \cdot)])$

Chapter 1

Introduction

1.1 Background

Variational inequalities (VIs) are a fundamental concept in mathematical analysis and optimization, introduced in the 1960s by Hartman and Stampacchia [31] as an important tool for analyzing boundary value problems of partial differential equations (PDEs) and optimal control problems. VIs provide a unified framework for representing various important concepts in applied mathematics such as nonlinear equation systems, complementarity problems, optimality conditions for optimization problems and network equilibrium problems. Thus, VIs have a wide range of applications in physics, economics, engineering sciences and so on [6, 21, 32, 59, 68].

Given a closed convex set $\Omega \subseteq \mathbb{R}^n$ and a continuous operator $H : \mathbb{R}^n \to \mathbb{R}^n$, the deterministic variational inequality (VI) is to find an $x^* \in \Omega$ such that

$$\langle H(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \Omega. \tag{1.1.1}$$

We denote problem (1.1.1) by VI(Ω, H), which includes the following important cases:

(i) if $H = \nabla f$ for some smooth function f, then problem (1.1.1) reduces to the first-order optimality condition for minimizing f over Ω;

(ii) if $\Omega = \mathbb{R}^n$, then problem (1.1.1) reduces to the nonlinear equation system:

$$H(x) = \mathbf{0};$$

(iii) if $\Omega = \mathbb{R}^n_+$, then problem (1.1.1) becomes the complementarity problem:

$$H(x) \geq \mathbf{0}, \quad x \geq \mathbf{0}, \quad \langle H(x), x \rangle = 0.$$

To address decision-making problems that involve future uncertainty, the stochastic version of VIs has been developed [37, 40, 45, 57, 78, 79]. Stochastic variational inequalities (SVIs) are a powerful mathematical framework that generalizes deterministic VIs to incorporate randomness and uncertainty. Let $(\Xi, \mathcal{G})$ be a measurable space and $T : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ a measurable (random) operator. $\xi : X \to \Xi$ is a random variable defined on a probability space $(X, \mathcal{F}, \mathbb{P})$. The stochastic variational inequality (SVI) is to find an $x^* \in \Omega$ such that

$$\langle \mathbb{E}[T(\xi, x^*)], x - x^* \rangle \geq 0, \quad \forall x \in \Omega. \tag{1.1.2}$$

We denote problem (1.1.2) by SVI $(\Omega, \mathbb{E}[T(\xi, \cdot)])$, which has many applications in practice [39, 48, 56].

Since calculating expectations requires multi-dimensional integration and the probability distribution of ξ may be unknown, existing numerical methods for the deterministic problem (1.1.1) are typically not directly applicable to the stochastic problem (1.1.2). To overcome this difficulty, stochastic approximation (SA) is a commonly used approximation technique.

In this thesis, we focus on designing accelerated algorithms for solving nonmonotone VIs (1.1.1) and (1.1.2).

1.2 Literature reviews

One of the most interesting topics in the study of VIs is the development of efficient and fast iterative algorithms for finding solutions. As a class of effective numerical

methods for solving VIs, projection methods have received a lot of attention from many researchers.

1.2.1 The projection methods for solving variational inequalities

The earliest projection method for solving VI(Ω, H) is the gradient projection (PG) method [28]

$$x_{k+1} = P_\Omega(x_k - tH(x_k)), \tag{1.2.1}$$

where P_Ω denotes the projection onto the set Ω and $t > 0$. To guarantee the convergence of the PG method, it is usually assumed that either H is strongly monotone and L-Lipschitz continuous on Ω[1], or H is cocoercive. In order to weaken the strong monotonicity of H, many projection methods have been proposed to solve monotone VIs [2, 14, 23, 15, 50, 51, 71, 72, 73]. These methods include using some acceleration techniques to solve monotone VIs. For example, an adaptive golden ratio algorithm was proposed in [2] for solving VIs where the operator is monotone and locally Lipschitz continuous. Specifically, when the feasible set is $\mathbb{R}^n$, the monotonicity assumption of the operator in [2] can be replaced by the existence of a weak Minty solution to VIs. Under these assumptions, convergence rates were provided for both the gap function and the residual function. In [23], the authors introduced an inertial projection and contraction algorithm and proved that, under the conditions that the operator is monotone and Lipschitz continuous, the sequence generated by the algorithm weakly converges to a solution of VIs in a Hilbert space.

Very recently, developing algorithms for different nonmonotone VIs has attracted great attention due to applications in machine learning [12, 46, 63, 64, 69, 77]. In this thesis, we focus on solving VIs with a pseudomonotone operator, which is a widely used class of nonmonotone VIs. A most commonly used algorithm in the literatures

[1] H is called L-Lipschitz continuous on Ω if $\|H(x) - H(y)\| \leq L\|x - y\|$ for any $x, y \in \Omega$.

for solving pseudomonotone VIs is the Korpelevich's extra gradient (EG) method [44]. The EG method was originally used to solve VI(Ω, H) with a monotone and L-Lipschitz continuous H, and was later extended by Facchinei and Pang [25] to solve the pseudomonotone VIs. The structure of the EG method proceeds as follows:

$$\begin{cases} x_{k+0.5} := P_{\Omega}(x_k - tH(x_k)), \\ x_{k+1} := P_{\Omega}(x_k - tH(x_{k+0.5})), \end{cases}$$

where $t > 0$. It is important to note that the convergence results for the EG method require H to be L-Lipschitz continuous on Ω and the stepsize t to satisfy $tL < 1$ [44]. Iusem [36] proposed a modified EG method with an updated stepsize to guarantee the efficiency of the proposed algorithm for VI(Ω, H), in which H is monotone and continuous. Recently, similar extensions have been developed not only for monotone operators but also for pseudomonotone operators [12, 63, 64]. However, the convergence rates are not mentioned in these works. Most results on convergence rates of EG methods for VIs are established based on the L-Lipschitz continuity of H, resulting in a sublinear rate of convergence for the best-iterate of the residual term. In particular, when H is monotone and L-Lipschitz continuous, we known that the EG method converges to a solution of VI(Ω, H) in terms of $\min_{0\leq k\leq N} \|x_k - x_{k+0.5}\|$ with a rate of $O(1/\sqrt{N})$ [60], which has been extended to the EG method for solving VIs with a pseudomonotone and L-Lipschitz continuous H in [25, Lemma 12.1.10].

One way to improve the EG method is to reduce the number of projections onto the feasible set. The EG method requires computing two projections onto the feasible set in each iteration. If the feasible set is simple enough, the projections are easy to compute. However, if the feasible set is a general closed convex set, two minimum distance problems must be solved to obtain the next iteration. For this reason, some scholars have improved the EG method by reducing the number of required projections.

In 2012, Censor, Gibali, and Reich introduced the subgradient extra gradient algorithm in [15]. In each iteration of the algorithm, a halfspace is constructed, replacing the second projection onto the feasible set with a projection onto this halfspace. This substitution is possible because the projection onto the constructed halfspace has a closed-form expression (see Lemma 2.6), resulting in lower computational complexity compared to the EG method. The subgradient extra gradient algorithm is structured as follows:

$$\begin{cases} x_{k+0.5} := P_{\Omega}(x_k - tH(x_k)), \\ B_k := \{x \in \mathbb{R}^n : \langle x_k - tH(x_k) - x_{k+0.5}, x - x_{k+0.5} \rangle \leq 0\}, \\ x_{k+1} := P_{B_k}(x_k - tH(x_{k+0.5})), \end{cases}$$

where $t > 0$. Under the assumptions that the operator H is monotone, L-Lipschitz continuous, and $tL < 1$, it was proved that the sequence generated by this algorithm weakly converges to a solution of $\mathrm{VI}(\Omega, H)$ in a Hilbert space. Subsequently, several scholars have extended the subgradient extra gradient algorithm in [14, 61, 65, 1, 18, 66, 74].

After that, the EG method has been intensively studied and extended in various ways [12, 14, 15, 22, 29, 35, 46, 49, 51, 63, 64, 69, 73, 75]. It is worth pointing out that besides the L-Lipschitz continuity of H on Ω, some other conditions on H are often required to guarantee the convergence of the EG method and its variants, such as Minty condition, quasimonotonicity and pseudomonotonicity.

Furthermore, research on nonmonotone VIs under Minty condition has been conducted. We say problem (1.1.1) satisfies the Minty condition if there exists an $x^* \in \Omega$ such that

$$\langle H(x), x - x^* \rangle \geq 0, \ \forall x \in \Omega. \tag{1.2.2}$$

Under the Minty condition, Ye and He in [77] introduced a double projection algorithm (DPA) with global convergence on the sequence, which requires computing

the projection onto the intersection of a finite number of halfspaces and the closed convex set Ω. Subsequently, Lei and He in [46] proposed a new extra gradient method (NEG) that does not involve adding halfspaces during the projection computation for solving this class of VIs under the same assumptions as in [77]. Then a new extra gradient type projection algorithm (NEGTP) was presented in [69] to solve a class of continuous quasimonotone VIs satisfying $H(x) \neq \mathbf{0}, \forall x \in \Omega$. All the algorithms in [46, 69, 77] have the global sequence convergence, but do not have the estimation of the convergence rate. In [76], Ye proved the global convergence of the sequence for the proposed algorithm under the Minty condition. However, in each iteration, the algorithm in [76] requires selecting the halfspace that has the largest distance from x_k to some generated halfspaces. As k increases, more information needs to be computed and stored. Approximation-based Regularized Extra-gradient method (ARE), a p^{th}-order ($p \geq 1$) algorithm, was proposed in [35] for solving monotone VIs with a convergence rate of $O(1/N^{\frac{p+1}{2}})$ on the gap function. In [34], it was stated that ARE also can solve the nonmonotone VIs satisfying the Minty condition with the convergence rate of $O(1/\sqrt{N})$ for the residual function and $O(1/N^{\frac{p}{2}})$ for the gap function. However, the algorithms in [35, 34] need the Lipschitz continuity of H and do not have the sequence convergence on the iterates. A summary of some main comparisons is provided in Table 1.1.

1.2.2 Stochastic approximation methods for solving stochastic variational inequalities

The SA method, first proposed by Robbins and Monro [57], was originally developed for stochastic root-finding problems. This approach involves substituting the exact mean operator $\mathbb{E}[T(\xi, \cdot)]$ with a random sample of T during the iteration process. The first SA-based projection method for SVIs is derived from the combination of the classical projection scheme and SA techniques, as introduced in [40]. This method

Methods	Assumptions	Sequence convergence	Convergence rate (residual function)
EG [25]	pseudomonotone Lipschitz continuous	$\surd$	$O(\frac{1}{\sqrt{N}})$
ARE [34] (the order $p = 1$)	Minty condition Lipschitz continuous	$\times$	$O(\frac{1}{\sqrt{N}})$
DPA [77]	Minty condition continuous	$\surd$	$\times$
NEG [46]	Minty condition continuous	$\surd$	$\times$
NEGTP [69]	Minty condition quasimonotone continuous $H(x) \neq 0, \forall x \in \Omega$	$\surd$	$\times$

Table 1.1: Summary of results on algorithms with global convergence to nonmonotone VIs

ensures convergence under the assumptions of Lipschitz continuity and strong monotonicity of $\mathbb{E}[T(\xi, \cdot)]$. Yousefian, Nedić and Shanbhag [78] extended the iterative scheme from [40] to solve Cartesian SVIs. The convergence conditions in [78] still require the operator $\mathbb{E}[T(\xi, \cdot)]$ to be strongly monotone and Lipschitz continuous.

Subsequently, some scholars weakened the strong monotonicity of the operator to monotonicity. Yousefian, Nedić and Shanbhag [79] proposed a regularized smoothing algorithm to solve monotone SVIs. Koshal, Nedić and Shanbhag [45] introduced a stochastic iterative Tikhonov regularization method for SVIs, which requires the operator $\mathbb{E}[T(\xi, \cdot)]$ to be monotone and Lipschitz continuous for convergence. Iusem, Jofré, Oliveira and Thompson [37] introduced a stochastic extra gradient method for

solving SVI $(\Omega, \mathbb{E}[T(\xi, \cdot)])$. The structure of the algorithm is as follows:

$$\begin{cases} x_{k+0.5} := P_\Omega \left(x_k - t \dfrac{1}{S_k} \displaystyle\sum_{j=1}^{S_k} T(\xi_j^k, x_k) \right), \\ x_{k+1} := P_\Omega \left(x_k - t \dfrac{1}{S_k} \displaystyle\sum_{j=1}^{S_k} T(\eta_j^k, x_{k+0.5}) \right), \end{cases}$$

where $t \in (0, \frac{1}{\sqrt{6}L})$, L is the Lipschitz constant of $\mathbb{E}[T(\xi, \cdot)]$, $\{S_k\}$ is the sample rate, and $\{\xi_j^k\}_{j=1}^{S_k}, \{\eta_j^k\}_{j=1}^{S_k}$ are random samples. For this algorithm to achieve almost sure convergence, the operator $\mathbb{E}[T(\xi, \cdot)]$ must be pseudomonotone.

Subsequently, Iusem, Jofré, Oliveira and Thompson [38] proposed a dynamic stochastic extra gradient method for SVIs, establishing the first robust SA algorithm with guaranteed convergence despite unknown Lipschitz constants. Based on this pioneering work, Zhang et al. [81] introduced an infeasible projection algorithm with line search for pseudomonotone SVIs, achieving the same iteration complexity but higher oracle complexity than [38]. However, unlike [38], each iteration in [81] requires only one projection and one sample family, significantly simplifying implementation. To reduce the computational cost, Yang et al. [74] proposed a variance-based subgradient extra gradient algorithm with line search for SVIs. The approach constructs a halfspace at each iteration and replaces the second projection onto the feasible set with a projection onto this halfspace. There are also several variants of variance-based extra gradient algorithms with line search for solving SVIs explored in references [42, 47, 70]. Cui, Sun and Zhang in [19] established theoretical foundations for the solvability of multistage pseudomonotone SVIs and proposed sufficient conditions for their elicitability, enabling the application of Rockafellar's progressive hedging algorithm. Additionally, Cui and Zhang in [20] proposed a localized progressive hedging algorithm for solving nonmonotone SVIs.

1.2.3 Anderson acceleration for fixed point problems

It is known that (1.1.1) is equivalent to the following fixed point problem [25]:

$$x = G(x) := P_\Omega(x - tH(x)), \tag{1.2.3}$$

where $t > 0$. Thus, the study of PG method in (1.2.1), which is a fixed point method, can help to improve the performance of the algorithms for solving VIs. Anderson acceleration is efficient to improve the convergence rate of fixed point methods, but the existing convergence analysis of Anderson acceleration requires the fixed point mapping G to be contractive and piecewise smooth.

Anderson acceleration was first proposed by Anderson in 1965 in the context of integral equations [4]. This technique aims to improve the convergence rate of fixed point iteration by utilizing the history of search directions. It is not necessary to compute the Jacobian of G, which allows it to perform effectively in various fields, including electronic structure computations [4, 16], machine learning [33], radiation diffusion and nuclear physics [3]. Anderson acceleration is formally described in the following algorithm, commonly referred to as Anderson(m).

Algorithm: Anderson(m)

Choose $x_0 \in \mathbb{R}^n$ and an integer $m > 0$. Set $x_1 = G(x_0)$ and $F_0 = G(x_0) - x_0$.
for $k = 1, 2, \ldots,$ **do**
set $F_k = G(x_k) - x_k$;
choose $m_k = \min\{m, k\}$;
solve

$$\min \left\| \sum_{j=0}^{m_k} \theta_j F_{k-m_k+j} \right\|, \qquad \text{s.t.} \quad \sum_{j=0}^{m_k} \theta_j = 1 \tag{1.2.4}$$

to find a solution $\{\theta_j^k : j = 0, \ldots, m_k\}$, and set

$$x_{k+1} = \sum_{j=0}^{m_k} \theta_j^k G(x_{k-m_k+j}).$$

end for

Even after a long period of use and attention, the first mathematical convergence result for Anderson acceleration had not been given until 2015 by Toth and Kelley [67]. They showed that when G is Lipschitz continuously differentiable and contractive, Anderson(m) has local r-linear convergence with r-factor $\hat{c} \in (c, 1)$, and Anderson(1) has q-linear convergence with q-factor c, where c is the contraction coefficient of the fixed point mapping. In 2019, Chen and Kelley [16] weakened the condition of G, proving that this conclusion can be obtained as long as G is a continuously differentiable operator. Additionally, Bian, Chen and Kelley [8] demonstrated the q-linear convergence of Anderson(1) for general nonsmooth fixed point problems in a Hilbert space, and r-linear convergence of Anderson(m) for a special nonsmooth operator. Then, Bian and Chen [7] proved that Anderson(1) is q-linear convergent for the composite max fixed point problem with a smaller q-factor than the existing q-factors. Zhang, O'Donoghue and Boyd [80] introduced a variant of Anderson acceleration that guaranteed global convergence for nonsmooth fixed point problems, but did not provide a convergence rate. Ouyang, Tao, Milzarek and Deng

[54] established a globalization strategy for Anderson acceleration incorporating a nonmonotone trust-region framework. They demonstrated that the algorithm has global convergence for a class of nonexpansive mappings and showed a local r-linear convergence for contractive mappings. Moreover, the local properties of Anderson acceleration with restarting were investigated in [53] in terms of function values when applied to a basic gradient scheme.

This thesis primarily investigates the case in which the operator H is pseudomonotone in problems (1.1.1) and (1.1.2). Due to the nonmonotonicity of H, the mapping G in (1.2.3) is not contractive, and may be even expansive, which means that we cannot directly use Anderson acceleration to solve the corresponding fixed point problem of $\mathrm{VI}(\Omega, H)$ in (1.2.3). Moreover, G is nonsmooth, because of the projection operator P_Ω. In this thesis, we introduce Anderson acceleration technique into the EG method to ensure the global sequence convergence of the algorithm and improve the convergence rate of the EG method.

1.3 Contribution of the thesis

The contributions of this thesis are summarized as follows.

- In the first part, we propose a new algorithm to solve pseudomonotone $\mathrm{VI}(\Omega, H)$ by combining Anderson(1) with the EG method. This algorithm uses the EG method with line search to guarantee the global convergence of the sequence and incorporates the Anderson acceleration strategy to enhance its convergence rate. We prove that the sequence generated by the proposed algorithm converges to a solution of pseudomonotone VIs from any initial point. This convergence is achieved without relying on the assumptions of Lipschitz continuity and contractive conditions, which are typically required for the convergence analysis of the EG algorithm and the Anderson acceleration algorithm

in existing literatures.

Under the assumption that H is locally Lipschitz continuous, the convergence rate of the proposed algorithm with respect to the residual function is no worse than that of the EG method. This assumption is weaker than the requirement of the EG method, which assumes that H is globally Lipschitz continuous.

Additionally, we propose an improved version of the algorithm by reducing the number of projections onto the feasible set. The improved algorithm is suitable for situations where projections onto the feasible set are difficult to compute. In these cases, the improved algorithm retains the same convergence properties as the original algorithm while improving its efficiency. Finally, numerical experiments are conducted to demonstrate the effectiveness and superiority of the proposed algorithm compared to the Anderson acceleration and the EG algorithms.

- In the second part, we develop an accelerated algorithm for solving pseudomonotone SVI $(\Omega, \mathbb{E}[T(\xi, \cdot)])$. This algorithm combines the EG method with Anderson acceleration and utilizes the SA method. The proposed stochastic algorithm does not require prior knowledge of the Lipschitz constant due to the implementation of a line search technique. We demonstrate that the sequence generated by the proposed stochastic algorithm converges almost surely to a solution of SVI $(\Omega, \mathbb{E}[T(\xi, \cdot)])$. Furthermore, we show that the proposed stochastic algorithm achieves a sublinear convergence rate of $O\left(N^{-1}\right)$ in terms of the mean residual function, as well as an optimal oracle complexity of $O(\epsilon^{-3})$. Finally, the numerical experiments ultimately demonstrate the superior performance of the proposed stochastic algorithm compared to the stochastic EG method. Furthermore, when applied to the portfolio selection with real data, the solution found by the proposed stochastic algorithm performs better than

the naive evenly weighted portfolio.

1.4 Organization of the thesis

The thesis is organized as follows.

- In Chapter 1, we provide an overview of the thesis background and conduct a review of the existing literatures on related topics.

- In Chapter 2, we establish the basic notation and summarize the key concepts and preliminary lemmas that underpin our analysis, including Opial's Lemma, which is essential for proving the main convergence results of this work.

- In Chapter 3, we propose an algorithm for solving pseudomonotone VIs, based on the Anderson acceleration and the EG methods. First, we prove the global convergence of the sequence generated by the proposed algorithm without relying on the Lipschitz continuity or contraction conditions of the operator. Then, under the condition that the operator is locally Lipschitz continuous, we provide the convergence rate concerning the residual function. Additionally, by reducing the number of projections onto the feasible set, we present an improved version of the proposed algorithm, enhancing its efficiency while maintaining its convergence properties. Finally, through numerical experiments, we demonstrate the effectiveness of the proposed algorithm, showing that its numerical performance is superior to both the Anderson acceleration and the EG methods.

- In Chapter 4, we present a stochastic version of the proposed algorithm using the SA method to solve pseudomonotone SVIs. The algorithm employs the line search to overcome the difficulty of knowing the Lipschitz constant of the operator in advance. We analyze the asymptotic convergence of the algorithm and

provide the convergence rate concerning the mean residual function and the oracle complexity. Finally, the proposed stochastic algorithm is shown to be effective compared to the stochastic EG algorithm through two illustrative examples. Moreover, the solution obtained by the proposed stochastic algorithm outperforms the naive evenly weighted portfolio on the portfolio selection problem with real data.

- In Chapter 5, we summarize the main results of the thesis and propose possible directions for future research.

Chapter 2

Basic Notation and Preliminaries

In this chapter, we introduce the basic notation and some fundamental concepts that will be used throughout the thesis.

2.1 Basic notation

Let $\|x\|$ denote the Euclidean norm of a vector $x \in \mathbb{R}^n$, and let $\|A\|$ and $\|A\|_F$ denote the ℓ_2-norm and the Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$, respectively. Given $\Omega \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, we use the notation $P_\Omega(x) := \arg\min_{y\in\Omega} \|y - x\|^2$ when Ω is closed. Let $\mathbb{N}$ be the set of all positive integers and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Let e denote the base of the natural logarithm, and $\mathbf{e} \in \mathbb{R}^n$ denote the vector of all ones, i.e., $\mathbf{e} = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$. For any $a \in \mathbb{R}$, $\lceil a \rceil$ represents the smallest integer greater than or equal to a. Given a random variable ξ and a σ-algebra $\mathcal{F}$, we denote the expectation of ξ by $\mathbb{E}[\xi]$ and the conditional expectation of ξ by $\mathbb{E}[\xi|\mathcal{F}]$. We use $[m] := \{1, 2, \cdots, m\}$ for $m \in \mathbb{N}$, and denote the σ-algebra generated by the random variables $\xi^1, \cdots, \xi^k$ as $\sigma(\xi^1, \cdots, \xi^k)$. Given the random variable ξ and $p \geq 1$, $|\xi|_p$ is the L^p-norm of ξ and $|\xi|\mathcal{F}|_p := (\mathbb{E}[|\xi|^p|\mathcal{F}])^{\frac{1}{p}}$ is the L^p-norm of ξ conditional to the σ-algebra $\mathcal{F}$. We write $\xi \in \mathcal{F}$ if ξ is $\mathcal{F}$-measurable, i.i.d. for ' independent identically distributed', and a.s. for 'almost surely'.

2.2 Preliminaries

Definition 2.1. *[43] Let Ω be a subset of $\mathbb{R}^n$ and $H : \mathbb{R}^n \to \mathbb{R}^n$ an operator. Then*

(i) H is said to be monotone on Ω, if for any $x, y \in \Omega$ it holds

$$\langle H(x) - H(y), x - y\rangle \geq 0;$$

(ii) H is said to be pseudomonotone on Ω, if for any $x, y \in \Omega$ it holds

$$\langle H(x), y - x\rangle \geq 0 \Rightarrow \langle H(y), y - x\rangle \geq 0.$$

Remark 2.1. *From Definition 2.1 and* (1.2.2)*, the following relation holds*

$$\textit{monotone} \Rightarrow \textit{pseudomonotone} \Rightarrow \textit{Minty condition.}$$

Lemma 2.1. *[27] For any $x \in \mathbb{R}^n$, the following statements hold.*

(i) $\|P_\Omega(x) - P_\Omega(y)\| \leq \|x - y\|, \quad \forall y \in \mathbb{R}^n;$

(ii) $\|P_\Omega(x) - P_\Omega(y)\|^2 \leq \langle P_\Omega(x) - P_\Omega(y), x - y\rangle, \quad \forall y \in \mathbb{R}^n;$

(iii) $\langle x - P_\Omega(x), y - P_\Omega(x)\rangle \leq 0, \quad \forall y \in \Omega.$

Lemma 2.2. *[22] For any $x \in \mathbb{R}^n$ and $t_1 \geq t_2 > 0$, the following inequalities hold:*

$$\frac{\|x - P_\Omega(x - t_1 H(x))\|}{t_1} \leq \frac{\|x - P_\Omega(x - t_2 H(x))\|}{t_2},$$

$$\|x - P_\Omega(x - t_2 H(x))\| \leq \|x - P_\Omega(x - t_1 H(x))\|.$$

Lemma 2.3. *[55]* [Opial's Lemma] *Let S be a nonempty subset of $\mathbb{R}^n$, and $\{x_k\}$ a sequence of elements in $\mathbb{R}^n$. Assume that*

(i) every sequential cluster point of $\{x_k\}$, as $k \to \infty$, belongs to S;

(ii) for every $z \in S$, $\lim_{k\to\infty} \|x_k - z\|$ exists.

Then the sequence $\{x_k\}$ converges as $k \to \infty$ to a point in S.

Lemma 2.4. *[29] Suppose that the mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ is continuous. Then, for all bounded sequences $\{x_k\}$, $\{y_k\} \subseteq \mathbb{R}^n$ satisfying $\lim_{k\to\infty} \|x_k - y_k\| = 0$, it holds that $\lim_{k\to\infty} \|H(x_k) - H(y_k)\| = 0$.*

Lemma 2.5. *[5] Let $\{a_k\}$ and $\{\varepsilon_k\}$ be real number sequences. Assume that $\{a_k\}$ is bounded from below, $\sum_{k=1}^{\infty} \varepsilon_k < \infty$ and*

$$a_{k+1} - a_k \leq \varepsilon_k$$

for every k. Then $\lim_{k\to\infty} a_k$ exists.

Lemma 2.6. *[10] Let v be a fixed vector and $B_{u,a} := \{x \in \mathbb{R}^n : \langle u, x\rangle \leq a\}$. Then*

$$P_{B_{u,a}}(v) = v - \max\left\{\frac{\langle u, v\rangle - a}{\|u\|^2}, 0\right\} u.$$

Definition 2.2. *[24] A σ-algebra $\mathcal{F}$ on a set X is a collection of subsets of X if it satisfies the following conditions:*

(i) $\emptyset \in \mathcal{F}$;

(ii) if $B \in \mathcal{F}$, then its complement B^c is also in $\mathcal{F}$;

(iii) if $B_1, B_2, \cdots$ is a countable collection of sets in $\mathcal{F}$, then $\cup_{n=1}^{\infty} B_n \in \mathcal{F}$.

Definition 2.3. *[24] If $\{x_k\}$ is a sequence with*

(i) $\mathbb{E}[|x_k|] < \infty$ for all k,

(ii) x_k is $\mathcal{F}_k$-measurable for each k,

(iii) $\mathbb{E}[x_{k+1}|\mathcal{F}_k] = x_k$ for all k,

then $\{x_k, \mathcal{F}_k\}$ is said to be a martingale.

Chapter 3

An Extra Gradient (EG) Anderson-Accelerated Algorithm for Pseudomonotone Variational Inequalities

In this chapter, we propose an accelerated algorithm for solving the pseudomonotone VI(Ω, H), based on the EG method and Anderson acceleration. In section 3.1, we present the framework and convergence analysis of the proposed algorithm. First, we prove that the proposed algorithm is well-defined. Then, we provide the global convergence of the sequence, followed by the convergence rate of the residual function, which is $O(N^{-1})$. Additionally, we present an improved version of the proposed algorithm and analyze its convergence properties in section 3.2. This is achieved by constructing a halfspace in each iteration and using the projection onto the halfspace instead of a second projection onto the feasible set. Since the halfspace projection has a closed-form solution, this algorithm is particularly useful when projections onto the feasible set are computationally expensive. It preserves the convergence properties of the original algorithm while improving efficiency. Finally, in section 3.3, we demonstrate the effectiveness of the proposed algorithm through five numerical examples.

3.1 EG-Anderson(1) algorithm and its convergence analysis

Reformulate the case where the operator in problem (1.1.1) is pseudomonotone: find an $x^* \in \Omega$ such that

$$\langle H(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \Omega, \tag{3.1.1}$$

where Ω is a closed convex set of $\mathbb{R}^n$ and $H : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous operator, and pseudomonotone on Ω, but not necessarily smooth or even Lipschitz continuous. Throughout this thesis, we denote the solution set of the problem (3.1.1) by $\mathrm{SOL}(\Omega, H)$, and assume $\mathrm{SOL}(\Omega, H) \neq \emptyset$.

We begin by introducing two operators that play a crucial role in the proposed algorithm, along with a lemma that will be used for its convergence analysis.

Define the following operators

$$G_t(x) := P_\Omega(x - tH(x)) \quad \text{and} \quad \tilde{G}_t(x) := P_\Omega(x - tH(G_t(x))),$$

where $t > 0$. Let

$$F_t(x) := G_t(x) - x \quad \text{and} \quad \tilde{F}_t(x) := \tilde{G}_t(x) - x.$$

Lemma 3.1. [25, Proposition 1.5.8, Exercise 1.8.29] *$x^* \in \mathrm{SOL}(\Omega, H)$ if and only if it is a fixed point of G_t, and if and only if it is a fixed point of $\tilde{G}_t$, where t can be any positive number.*

3.1.1 Framework of the EG-Anderson(1) algorithm

The proposed algorithm, called the EG-Anderson(1) algorithm, is presented in Algorithm 1. In this algorithm, Step 2 incorporates the line search framework introduced in [12], which is designed to adaptively select the step size and thereby improve the robustness and efficiency of the iterative process.

Algorithm 1: EG-Anderson(1)

Initialization: Choose $x_0 \in \Omega$, $\nu \geq 0, \gamma > 0$, $\tau > \frac{1}{2}$, $\rho, \mu \in (0,1)$ and $\sigma_0 = 1$. Give a sufficiently large $M > 0$.

for $k = 0, 1, 2, \ldots,$ **do**

Step 1: Compute $F_\gamma(x_k) = P_\Omega(x_k - \gamma H(x_k)) - x_k$.

If $F_\gamma(x_k) = \mathbf{0}$, then stop.

Otherwise, let $t_k = \gamma$ and go to **Step 2**.

Step 2: Compute

$$y_{k+0.5} = P_\Omega(x_k - t_k H(x_k)) \quad \text{and} \quad y_{k+1} = P_\Omega(x_k - t_k H(y_{k+0.5})).$$

If

$$\begin{aligned} & t_k \langle H(y_{k+0.5}) - H(x_k), y_{k+0.5} - y_{k+1} \rangle \\ & \leq \frac{\mu}{2} \left(\|x_k - y_{k+0.5}\|^2 + \|y_{k+0.5} - y_{k+1}\|^2 \right), \end{aligned} \tag{3.1.2}$$

go to **Step 3**.

Otherwise, set $t_k = \rho t_k$ and repeat **Step 2**.

Step 3: Compute $F_{t_k}(x_k) = y_{k+0.5} - x_k$ and $\tilde{F}_{t_k}(x_k) = y_{k+1} - x_k$.

If $\|\tilde{F}_{t_k}(x_k)\| < \min\{\|F_{t_k}(x_k)\|, \nu \sigma_k^{-\tau}\}$, set

$$\alpha_k = \frac{\langle \tilde{F}_{t_k}(x_k), \tilde{F}_{t_k}(x_k) - F_{t_k}(x_k) \rangle}{\|\tilde{F}_{t_k}(x_k) - F_{t_k}(x_k)\|^2}. \tag{3.1.3}$$

Otherwise, set $\alpha_k = M + 1$.

Step 4: If $|\alpha_k| \leq M$, set

$$x_{k+1} = \alpha_k x_k + (1 - \alpha_k) y_{k+1}, \quad \sigma_{k+1} = \sigma_k + 1. \tag{3.1.4}$$

Otherwise, set

$$x_{k+1} = y_{k+1}, \quad \sigma_{k+1} = \sigma_k. \tag{3.1.5}$$

end for

From the notations and definitions for $y_{k+0.5}$ and y_{k+1} in the EG-Anderson(1) algorithm, we find that

$$y_{k+0.5} = G_{t_k}(x_k) \quad \text{and} \quad y_{k+1} = \tilde{G}_{t_k}(x_k);$$

$$y_{k+0.5} - x_k = F_{t_k}(x_k) \quad \text{and} \quad y_{k+1} - x_k = \tilde{F}_{t_k}(x_k).$$

3.1.2 Convergence analysis

In this subsection, we will analyze the convergence properties of the EG-Anderson(1) algorithm, including the global convergence of the sequence and the convergence rate evaluated by the residual function. In order to categorize the iteration counts, we divide them into two subsets:

$$K_{AA} = \{k_0, k_1, \cdots\} \quad \text{and} \quad K_{EG} = \{l_0, l_1, \cdots\},$$

where K_{AA} consists of iterations setting by (3.1.4) and K_{EG} includes the remaining iterations setting by (3.1.5).

If the EG-Anderson(1) algorithm is terminated in finite iterations, then the final output point is a solution of the problem (3.1.1). Therefore, in the following analysis we assume that the EG-Anderson(1) algorithm loops infinitely.

Remark 3.1. *For $k_i \in K_{AA}$, note that $\|\tilde{F}_{t_{k_i}}(x_{k_i}) - F_{t_{k_i}}(x_{k_i})\| \neq 0$ due to $\|\tilde{F}_{t_{k_i}}(x_{k_i})\| < \|F_{t_{k_i}}(x_{k_i})\|$, thus α_{k_i} is well-defined. Moreover, α_{k_i} is the optimal solution of*

$$\min_{\alpha \in \mathbb{R}} \left\| \alpha F_{t_{k_i}}(x_{k_i}) + (1-\alpha)\tilde{F}_{t_{k_i}}(x_{k_i}) \right\|.$$

We start the convergence analysis of the EG-Anderson(1) algorithm by proving that (3.1.2) terminates after a finite number of loops.

Lemma 3.2. *The EG-Anderson(1) algorithm is well-defined.*

Proof We will show that the EG-Anderson(1) algorithm is well-defined by proving that for every k there exists t_k satisfying (3.1.2) when $x_k \notin \mathrm{SOL}(\Omega, H)$.

From the updated form of t_k in the EG-Anderson(1) algorithm, it can be reformulated as $t_k = \gamma\rho^{m_k}$, where m_k is the smallest nonnegative integer m satisfying

$$\begin{aligned} &\gamma\rho^m \left\langle H\left(y_{k+0.5}^{(m)}\right) - H\left(x_k\right), y_{k+0.5}^{(m)} - y_{k+1}^{(m)} \right\rangle \\ &\leq \frac{\mu}{2}\left(\left\| x_k - y_{k+0.5}^{(m)} \right\|^2 + \left\| y_{k+0.5}^{(m)} - y_{k+1}^{(m)} \right\|^2 \right), \end{aligned} \tag{3.1.6}$$

where $y_{k+0.5}^{(m)} := P_\Omega\left(x_k - \gamma\rho^m H\left(x_k\right)\right)$ and $y_{k+1}^{(m)} := P_\Omega\left(x_k - \gamma\rho^m H\left(y_{k+0.5}^{(m)}\right)\right)$.

If there exists a nonnegative integer $\bar{m}$ such that $y_{k+0.5}^{(\bar{m})} = y_{k+1}^{(\bar{m})}$, then there must exist an integer m_k between 0 and $\bar{m}$ such that (3.1.6) holds. We consider the situation $y_{k+0.5}^{(m)} \neq y_{k+1}^{(m)}$ for any nonnegative integer m and assume the contrary that for all m we have

$$\gamma\rho^m \left\langle H\left(y_{k+0.5}^{(m)}\right) - H(x_k), y_{k+0.5}^{(m)} - y_{k+1}^{(m)} \right\rangle > \frac{\mu}{2}\left(\left\|x_k - y_{k+0.5}^{(m)}\right\|^2 + \left\|y_{k+0.5}^{(m)} - y_{k+1}^{(m)}\right\|^2\right). \tag{3.1.7}$$

On one hand, by Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} &\gamma\rho^m \left\langle H\left(y_{k+0.5}^{(m)}\right) - H\left(x_k\right), y_{k+0.5}^{(m)} - y_{k+1}^{(m)} \right\rangle \\ \leq &\gamma\rho^m \left\|H\left(y_{k+0.5}^{(m)}\right) - H(x_k)\right\| \left\|y_{k+0.5}^{(m)} - y_{k+1}^{(m)}\right\|. \end{aligned} \tag{3.1.8}$$

On the other hand, we also find

$$\left\|x_k - y_{k+0.5}^{(m)}\right\|^2 + \left\|y_{k+0.5}^{(m)} - y_{k+1}^{(m)}\right\|^2 \geq 2\left\|x_k - y_{k+0.5}^{(m)}\right\| \left\|y_{k+0.5}^{(m)} - y_{k+1}^{(m)}\right\|. \tag{3.1.9}$$

Combining (3.1.7) with (3.1.8) and (3.1.9), we deduce that

$$\frac{\left\|x_k - y_{k+0.5}^{(m)}\right\|}{\gamma\rho^m} \leq \frac{1}{\mu}\left\|H\left(y_{k+0.5}^{(m)}\right) - H(x_k)\right\|. \tag{3.1.10}$$

Since $x_k \notin \mathrm{SOL}(\Omega, H)$, we discuss the following two cases.

(i) If $x_k \in \Omega$, from the definition of $y_{k+0.5}^{(m)}$ and the continuity of P_Ω, we have

$$\lim_{m\to\infty}\left\|x_k - y_{k+0.5}^{(m)}\right\| = 0.$$

In view of the continuity of H, we get $\lim_{m\to\infty}\left\|H(x_k) - H\left(y_{k+0.5}^{(m)}\right)\right\| = 0$. This together with (3.1.10) yields

$$\lim_{m\to\infty}\frac{\left\|x_k - y_{k+0.5}^{(m)}\right\|}{\gamma\rho^m} = 0. \tag{3.1.11}$$

By the definition of $y_{k+0.5}^{(m)}$ and using Lemma 2.1-(iii), we get

$$\left\langle y_{k+0.5}^{(m)} - x_k + \gamma\rho^m H(x_k), x - y_{k+0.5}^{(m)} \right\rangle \geq 0, \quad \forall x \in \Omega,$$

which implies

$$\left\langle \frac{y_{k+0.5}^{(m)} - x_k}{\gamma\rho^m} + H(x_k), x - y_{k+0.5}^{(m)} \right\rangle \geq 0, \quad \forall x \in \Omega. \tag{3.1.12}$$

Taking the limit $m \to \infty$ in (3.1.12) and using (3.1.11) and $\lim_{m\to\infty} y_{k+0.5}^{(m)} = x_k$, we obtain $\langle H(x_k), x - x_k\rangle \geq 0, \forall x \in \Omega$. It can be deduced that $x_k \in \mathrm{SOL}(\Omega, H)$ and this leads to a contraction.

(ii) If $x_k \notin \Omega$, we can conclude that

$$\lim_{m\to\infty} \left\| x_k - y_{k+0.5}^{(m)} \right\| = \|x_k - P_\Omega(x_k)\| > 0, \tag{3.1.13}$$

and

$$\lim_{m\to\infty} \gamma\rho^m \left\| H\left(y_{k+0.5}^{(m)}\right) - H(x_k) \right\| = 0. \tag{3.1.14}$$

Rearranging the terms in (3.1.10), we find

$$\left\| x_k - y_{k+0.5}^{(m)} \right\| \leq \frac{1}{\mu}\gamma\rho^m \left(\left\| H\left(y_{k+0.5}^{(m)}\right) - H(x_k) \right\| \right).$$

Taking the limit $m \to \infty$ in the above inequality, we can find a contradiction with (3.1.13) and (3.1.14). Hence, the proof is fully established.

□

The next two lemmas are instrumental in establishing the key findings of this section.

Lemma 3.3. *Let $\{x_k\}$ be the sequence generated by the EG-Anderson(1) algorithm. Then we have*

$$\left(1-\sqrt{\frac{\mu}{2-\mu}}\right)\|F_{t_k}(x_k)\| \le \|\tilde{F}_{t_k}(x_k)\| \le \left(1+\sqrt{\frac{\mu}{2-\mu}}\right)\|F_{t_k}(x_k)\|.$$

Proof For any k, by the condition of t_k in (3.1.2) and Lemma 2.1-(ii), we obtain

$$\begin{aligned}\|y_{k+0.5}-y_{k+1}\|^2 &= \|P_\Omega(x_k - t_k H(x_k)) - P_\Omega(x_k - t_k H(y_{k+0.5}))\|^2 \\ &\le \langle y_{k+0.5}-y_{k+1}, t_k\left(H(y_{k+0.5})-H(x_k)\right)\rangle \\ &\le \frac{\mu}{2}\|x_k-y_{k+0.5}\|^2 + \frac{\mu}{2}\|y_{k+0.5}-y_{k+1}\|^2,\end{aligned}$$

which implies

$$\|y_{k+1}-y_{k+0.5}\|^2 \le \frac{\mu}{2-\mu}\|x_k-y_{k+0.5}\|^2. \tag{3.1.15}$$

Since $\mu \in (0,1)$, then $\frac{\mu}{2-\mu} \in (0,1)$. From the triangle inequality and (3.1.15), we have

$$\begin{aligned}\|\tilde{F}_{t_k}(x_k)\| = \|y_{k+1}-x_k\| &\ge \|y_{k+0.5}-x_k\| - \|y_{k+1}-y_{k+0.5}\| \\ &\ge \|y_{k+0.5}-x_k\| - \sqrt{\frac{\mu}{2-\mu}}\|y_{k+0.5}-x_k\| \\ &= \left(1-\sqrt{\frac{\mu}{2-\mu}}\right)\|F_{t_k}(x_k)\|\end{aligned}$$

and

$$\begin{aligned}\|\tilde{F}_{t_k}(x_k)\| = \|y_{k+1}-x_k\| &\le \|y_{k+0.5}-x_k\| + \|y_{k+1}-y_{k+0.5}\| \\ &\le \left(1+\sqrt{\frac{\mu}{2-\mu}}\right)\|F_{t_k}(x_k)\|.\end{aligned}$$

The proof is completed. □

Lemma 3.4. *Let $\{x_k\}$, $\{y_{k+0.5}\}$ and $\{y_{k+1}\}$ be the sequences generated by the EG-Anderson(1) algorithm and $x^* \in \mathrm{SOL}(\Omega, H)$. For every k, it holds that*

$$\|y_{k+1}-x^*\|^2 \le \|x_k-x^*\|^2 - (1-\mu)\|y_{k+0.5}-x_k\|^2 - (1-\mu)\|y_{k+1}-y_{k+0.5}\|^2.$$

Proof In view of the pseudomonotonicity of H and $x^* \in \mathrm{SOL}(\Omega, H)$, we deduce that $\langle t_k H(y_{k+0.5}), x^* - y_{k+0.5}\rangle \leq 0$, which gives

$$\langle t_k H(y_{k+0.5}), x^* - y_{k+1}\rangle \leq \langle t_k H(y_{k+0.5}), y_{k+0.5} - y_{k+1}\rangle. \tag{3.1.16}$$

By the definition of $y_{k+0.5}$, Lemma 2.1-(iii) and (3.1.2), we obtain

$$\begin{aligned}
&\langle x_k - t_k H(y_{k+0.5}) - y_{k+0.5}, y_{k+1} - y_{k+0.5}\rangle \\
&= \langle x_k - t_k H(x_k) - y_{k+0.5}, y_{k+1} - y_{k+0.5}\rangle + t_k\langle H(x_k) - H(y_{k+0.5}), y_{k+1} - y_{k+0.5}\rangle \\
&\overset{(3.1.2)}{\leq} \langle x_k - t_k H(x_k) - P_\Omega(x_k - t_k H(x_k)), y_{k+1} - P_\Omega(x_k - t_k H(x_k))\rangle \\
&\quad + \frac{\mu}{2}\|x_k - y_{k+0.5}\|^2 + \frac{\mu}{2}\|y_{k+0.5} - y_{k+1}\|^2 \\
&\leq \frac{\mu}{2}\|x_k - y_{k+0.5}\|^2 + \frac{\mu}{2}\|y_{k+0.5} - y_{k+1}\|^2.
\end{aligned} \tag{3.1.17}$$

Let $z_k = x_k - t_k H(y_{k+0.5})$ for simplicity. Based on the definition of y_{k+1}, Lemma 2.1-(iii), (3.1.16) and (3.1.17), we conclude that

$$\begin{aligned}
\|y_{k+1} - x^*\|^2 &= \|P_\Omega(z_k) - x^*\|^2 \\
&= \|z_k - x^*\|^2 - \|z_k - P_\Omega(z_k)\|^2 + 2\langle P_\Omega(z_k) - z_k, P_\Omega(z_k) - x^*\rangle \\
&\leq \|z_k - x^*\|^2 - \|z_k - P_\Omega(z_k)\|^2 \\
&= \|x_k - t_k H(y_{k+0.5}) - x^*\|^2 - \|x_k - t_k H(y_{k+0.5}) - y_{k+1}\|^2 \\
&\leq \|x_k - x^*\|^2 - \|x_k - y_{k+1}\|^2 + 2\langle t_k H(y_{k+0.5}), x^* - y_{k+1}\rangle \\
&\overset{(3.1.16)}{\leq} \|x_k - x^*\|^2 - \|x_k - y_{k+1}\|^2 + 2\langle t_k H(y_{k+0.5}), y_{k+0.5} - y_{k+1}\rangle \\
&= \|x_k - x^*\|^2 - \|x_k - y_{k+0.5}\|^2 - \|y_{k+0.5} - y_{k+1}\|^2 \\
&\quad + 2\langle x_k - t_k H(y_{k+0.5}) - y_{k+0.5}, y_{k+1} - y_{k+0.5}\rangle \\
&\overset{(3.1.17)}{\leq} \|x_k - x^*\|^2 - (1-\mu)\|x_k - y_{k+0.5}\|^2 - (1-\mu)\|y_{k+0.5} - y_{k+1}\|^2.
\end{aligned} \tag{3.1.18}$$

□

Now, utilizing the Opial's Lemma, we can state and prove our main convergence result in what follows.

Theorem 3.1. *Let $\{x_k\}$ be the sequence generated by the EG-Anderson(1) algorithm. Then the sequence $\{x_k\}$ converges to a solution of the problem* (3.1.1).

Proof We will prove this theorem from three steps.

Step 1: $\{x_k\}$ is bounded.

Let x^* be a solution of problem $\mathrm{VI}(\Omega, H)$. If $k_i \in K_{AA}$, let $\beta_{k_i} := 1 - \alpha_{k_i}$ and by the definition of x_{k_i+1} in (3.1.4), we know that

$$\begin{aligned}
&\|x_{k_i+1} - x^*\|^2 \\
&= \|\alpha_{k_i} x_{k_i} + \beta_{k_i} y_{k_i+1} - x^*\|^2 \\
&= \alpha_{k_i}^2 \|x_{k_i} - x^*\|^2 + \beta_{k_i}^2 \|y_{k_i+1} - x^*\|^2 + 2\alpha_{k_i}\beta_{k_i}\langle x_{k_i} - x^*, y_{k_i+1} - x^*\rangle \\
&= \alpha_{k_i}^2 \|x_{k_i} - x^*\|^2 + \beta_{k_i}^2 \|y_{k_i+1} - x^*\|^2 \\
&\quad + 2\alpha_{k_i}\beta_{k_i}\left(\frac{1}{2}\|x_{k_i} - x^*\|^2 + \frac{1}{2}\|y_{k_i+1} - x^*\|^2 - \frac{1}{2}\|x_{k_i} - y_{k_i+1}\|^2\right) \\
&= \alpha_{k_i}\|x_{k_i} - x^*\|^2 + \beta_{k_i}\|y_{k_i+1} - x^*\|^2 - \alpha_{k_i}\beta_{k_i}\|x_{k_i} - y_{k_i+1}\|^2.
\end{aligned} \tag{3.1.19}$$

From Lemma 3.4, we can obtain

$$\|y_{k_i+1} - x^*\|^2 \le \|x_{k_i} - x^*\|^2 - (1-\mu)\|y_{k_i+0.5} - x_{k_i}\|^2 - (1-\mu)\|y_{k_i+1} - y_{k_i+0.5}\|^2. \tag{3.1.20}$$

Since $\|\tilde{F}_{t_{k_i}}(x_{k_i})\| < \|F_{t_{k_i}}(x_{k_i})\|$ when $k_i \in K_{AA}$, then

$$\langle F_{t_{k_i}}(x_{k_i}), \tilde{F}_{t_{k_i}}(x_{k_i})\rangle < \|F_{t_{k_i}}(x_{k_i})\|^2,$$

which implies

$$\langle \tilde{F}_{t_{k_i}}(x_{k_i}), \tilde{F}_{t_{k_i}}(x_{k_i}) - F_{t_{k_i}}(x_{k_i})\rangle < \|\tilde{F}_{t_{k_i}}(x_{k_i}) - F_{t_{k_i}}(x_{k_i})\|^2.$$

By (3.1.3), we have $\alpha_{k_i} < 1$. Thus $\beta_{k_i} = 1 - \alpha_{k_i} > 0$.

Introducing (3.1.20) into (3.1.19), we deduce that

$$\begin{aligned}\|x_{k_i+1}-x^*\|^2 \le \|x_{k_i}-x^*\|^2 &-(1-\mu)\beta_{k_i}\|y_{k_i+0.5}-x_{k_i}\|^2\\ &-(1-\mu)\beta_{k_i}\|y_{k_i+0.5}-y_{k_i+1}\|^2-\alpha_{k_i}\beta_{k_i}\|x_{k_i}-y_{k_i+1}\|^2.\end{aligned}\tag{3.1.21}$$

In view of $|\alpha_{k_i}| \le M$ and $\beta_{k_i} = 1-\alpha_{k_i}$, we get $|\beta_{k_i}| \le M+1$. This together with (3.1.21), $\mu \in (0,1)$, $\beta_{k_i} > 0$ and $\|x_{k_i} - y_{k_i+1}\| = \|\tilde{F}_{t_{k_i}}(x_{k_i})\| \le \nu\sigma_{k_i}^{-\tau} = \nu(\sigma_0+i)^{-\tau} = \nu(1+i)^{-\tau}$, we deduce that

$$\begin{aligned}\|x_{k_i+1}-x^*\|^2 &\le \|x_{k_i}-x^*\|^2 + |\alpha_{k_i}||\beta_{k_i}|\|\tilde{F}_{t_{k_i}}(x_{k_i})\|^2\\ &\le \|x_{k_i}-x^*\|^2+\varepsilon_{k_i},\end{aligned}\tag{3.1.22}$$

where $\varepsilon_{k_i} := M(M+1)\nu^2(1+i)^{-2\tau}$.

If $l_j \in K_{EG}$, by Lemma 3.4 and $\mu \in (0,1)$, we conclude that

$$\begin{aligned}\|x_{l_j+1}-x^*\|^2 &= \|y_{l_j+1}-x^*\|^2\\ &\le \|x_{l_j}-x^*\|^2-(1-\mu)\|y_{l_j+0.5}-x_{l_j}\|^2-(1-\mu)\|y_{l_j+1}-y_{l_j+0.5}\|^2\\ &\le \|x_{l_j}-x^*\|^2.\end{aligned}\tag{3.1.23}$$

By defining $\varepsilon_{l_j} = 0$ and combining (3.1.22) and (3.1.23), for every k, we find

$$\|x_{k+1}-x^*\|^2 \le \|x_k-x^*\|^2+\varepsilon_k,\tag{3.1.24}$$

where $\sum_{k=0}^{\infty}\varepsilon_k = \sum_{i=0}^{\infty}\varepsilon_{k_i} < \infty$ because of $\tau > \frac{1}{2}$. From (3.1.24), we further have

$$\|x_k-x^*\|^2 \le \|x_0-x^*\|^2+\sum_{k=0}^{\infty}\varepsilon_k < \infty.$$

Hence $\{x_k\}$ is bounded.

Step 2: any cluster point of $\{x_k\}$ belongs to $\mathrm{SOL}(\Omega, H)$, i.e., a solution of the problem (3.1.1).

By the boundedness of $\{x_k\}$, it has at least one cluster point, denoted by $\bar{x}$ with convergence subsequence $\{x_{p_j}\}$ satisfying $\lim_{j\to\infty} x_{p_j} = \bar{x}$. Next, we will prove $\bar{x} \in \mathrm{SOL}(\Omega, H)$ from the following two steps.

Step 2.1: $\lim_{k\to\infty}\|F_{t_k}(x_k)\| = 0$.

We know that $K_{AA} \cup K_{EG}$ is infinite. The following proof is divided into two cases for discussion.

(i) We first consider the case that both K_{AA} and K_{EG} are infinite. Rearranging the terms in (3.1.23) and using $\mu \in (0,1)$, we infer that

$$(1-\mu)\|F_{t_{l_j}}(x_{l_j})\|^2 = (1-\mu)\|y_{l_j+0.5}-x_{l_j}\|^2 \leq \|x_{l_j}-x^*\|^2-\|x_{l_j+1}-x^*\|^2. \quad (3.1.25)$$

Combining (3.1.22) and (3.1.25), we deduce that

$$\sum_{j=0}^{\infty}\|F_{t_{l_j}}(x_{l_j})\|^2 \leq \frac{1}{1-\mu}\|x_0-x^*\|^2 + \frac{1}{1-\mu}\sum_{i=0}^{\infty}\varepsilon_{k_i} < \infty. \quad (3.1.26)$$

Moreover, we know $\|\tilde{F}_{t_{k_i}}(x_{k_i})\| \leq \nu(1+i)^{-\tau}$ when $k_i \in K_{AA}$. Together with Lemma 3.3, we have

$$\|F_{t_{k_i}}(x_{k_i})\|^2 \leq \left(1-\sqrt{\frac{\mu}{2-\mu}}\right)^{-2}\|\tilde{F}_{t_{k_i}}(x_{k_i})\|^2 \leq \left(1-\sqrt{\frac{\mu}{2-\mu}}\right)^{-2}\nu^2(1+i)^{-2\tau}.$$

Thus

$$\sum_{i=0}^{\infty}\|F_{t_{k_i}}(x_{k_i})\|^2 \leq \left(1-\sqrt{\frac{\mu}{2-\mu}}\right)^{-2}\nu^2\sum_{i=0}^{\infty}(1+i)^{-2\tau} < \infty. \quad (3.1.27)$$

Since $\tau > \frac{1}{2}$, there exists a $C_0 > 0$ such that $\sum_{i=0}^{\infty}(1+i)^{-2\tau} \leq C_0$. Combining (3.1.26) and (3.1.27), we deduce that

$$\sum_{k=0}^{\infty}\|F_{t_k}(x_k)\|^2 = \sum_{i=0}^{\infty}\|F_{t_{k_i}}(x_{k_i})\|^2 + \sum_{j=0}^{\infty}\|F_{t_{l_j}}(x_{l_j})\|^2 \leq C, \quad (3.1.28)$$

where $C := \left(\left(1-\sqrt{\frac{\mu}{2-\mu}}\right)^{-2} + \frac{1}{1-\mu}M(M+1)\right)\nu^2C_0 + \frac{1}{1-\mu}\|x_0-x^*\|^2$. Hence we conclude that $\lim_{k\to\infty}\|F_{t_k}(x_k)\| = 0$.

(ii) We then consider the case that either K_{AA} or K_{EG} is finite. In this situation, the proof becomes simpler, as we only need to use either (3.1.26) or (3.1.27) to obtain $\sum_{k=0}^{\infty} \|F_{t_k}(x_k)\|^2 < \infty$.

Therefore, we have $\lim_{k\to\infty} \|F_{t_k}(x_k)\| = 0$.

Step 2.2: $\lim_{k\to\infty} \|F_\gamma(x_k)\| = 0$.

Let $\tilde{y}_{k+0.5} := P_\Omega(x_k - t_k\rho^{-1}H(x_k))$. Applying Lemma 2.2 and $t_k\rho^{-1} > t_k$, we get

$$\|x_k - \tilde{y}_{k+0.5}\| \le \rho^{-1}\|x_k - y_{k+0.5}\| = \rho^{-1}\|F_{t_k}(x_k)\|, \tag{3.1.29}$$

which implies $\lim_{k\to\infty} \|x_k - \tilde{y}_{k+0.5}\| = 0$.

Below, we will estimate $\frac{1}{t_k}\|x_k - \tilde{y}_{k+0.5}\|$ by dividing it into two cases based on the value of t_k.

(i) If $t_k = \gamma$, we have

$$\frac{1}{t_k}\|x_k - \tilde{y}_{k+0.5}\| = \frac{1}{\gamma}\|x_k - \tilde{y}_{k+0.5}\|. \tag{3.1.30}$$

(ii) If $t_k < \gamma$, let $\tilde{y}_{k+1} := P_\Omega(x_k - t_k\rho^{-1}H(\tilde{y}_{k+0.5}))$. From the condition of t_k in (3.1.2), we know that $t_k\rho^{-1}$ satisfies

$$t_k\rho^{-1}\left\langle H(\tilde{y}_{k+0.5}) - H(x_k), \tilde{y}_{k+0.5} - \tilde{y}_{k+1}\right\rangle > \frac{\mu}{2}\left(\|x_k - \tilde{y}_{k+0.5}\|^2 + \|\tilde{y}_{k+0.5} - \tilde{y}_{k+1}\|^2\right), \tag{3.1.31}$$

which implies $\tilde{y}_{k+0.5} \neq \tilde{y}_{k+1}$. Then for (3.1.31), based on a similar estimation as for (3.1.7), we can conclude that

$$\frac{1}{t_k}\|x_k - \tilde{y}_{k+0.5}\| < \mu^{-1}\rho^{-1}\|H(x_k) - H(\tilde{y}_{k+0.5})\|. \tag{3.1.32}$$

Combining (3.1.30) and (3.1.32), we can obtain

$$\frac{1}{t_k}\|x_k - \tilde{y}_{k+0.5}\| \le \max\left\{\frac{1}{\gamma}\|x_k - \tilde{y}_{k+0.5}\|, \mu^{-1}\rho^{-1}\|H(x_k) - H(\tilde{y}_{k+0.5})\|\right\}. \tag{3.1.33}$$

Since $\{x_k\}$ is bounded and $\lim_{k\to\infty}\|x_k-\tilde{y}_{k+0.5}\|=0$, we find $\{\tilde{y}_{k+0.5}\}$ is bounded. Combining the continuity of H and Lemma 2.4, we conclude that

$$\lim_{k\to\infty}\|H(x_k)-H(\tilde{y}_{k+0.5})\|=0.$$

This together with $\lim_{k\to\infty}\|x_k-\tilde{y}_{k+0.5}\|=0$ and (3.1.33) yields

$$\lim_{k\to\infty}\frac{1}{t_k}\|x_k-\tilde{y}_{k+0.5}\|=0.$$

Again by Lemma 2.2 and $t_k\rho^{-1}>t_k$, we get

$$\|x_k-\tilde{y}_{k+0.5}\|\geq\|x_k-y_{k+0.5}\|. \tag{3.1.34}$$

Then, we obtain

$$\lim_{k\to\infty}\frac{1}{t_k}\|x_k-y_{k+0.5}\|\leq\lim_{k\to\infty}\frac{1}{t_k}\|x_k-\tilde{y}_{k+0.5}\|=0,$$

which means $\lim_{k\to\infty}\frac{1}{t_k}\|F_{t_k}(x_k)\|=\lim_{k\to\infty}\frac{1}{t_k}\|x_k-y_{k+0.5}\|=0$. Hence, we conclude that $\lim_{k\to\infty}\frac{1}{t_k}\|F_{t_k}(x_k)\|=0$.

Recalling Lemma 2.2 and $t_k\leq\gamma$, we have $\frac{1}{t_k}\|F_{t_k}(x_k)\|\geq\frac{1}{\gamma}\|F_\gamma(x_k)\|$ and $\|F_{t_k}(x_k)\|\leq\|F_\gamma(x_k)\|$. Then we conclude that

$$\|F_{t_k}(x_k)\|\leq\|F_\gamma(x_k)\|\leq\frac{\gamma}{t_k}\|F_{t_k}(x_k)\|, \tag{3.1.35}$$

which gives

$$\lim_{k\to\infty}\|F_\gamma(x_k)\|=0.$$

Together with $\lim_{j\to\infty}x_{p_j}=\bar{x}$ and Lemma 3.1, we deduce that $\bar{x}\in\mathrm{SOL}(\Omega,H)$.

Step 3: $\{x_k\}$ is convergent to a solution of the problem (3.1.1).

Since $\|x_k-x^*\|^2$ is nonnegative, $\sum_{k=0}^{\infty}\varepsilon_k<\infty$ and (3.1.24) holds, by applying Lemma 2.5 with $a_k=\|x_k-x^*\|^2$, $\lim_{k\to\infty}\|x_k-x^*\|^2$ exists for any $x^*\in\mathrm{SOL}(\Omega,H)$. Together this with Step 2, the proof is completed by Lemma 2.3. □

If H is also locally Lipschitz continuous at any solution of the problem (3.1.1), we derive the following conclusion about the convergence rate on the residual function.

Theorem 3.2. *(Best-iterate convergence rate) Suppose that H is locally Lipschitz continuous at any solution of the problem* (3.1.1). *Let $\{x_k\}$ be the sequence generated by the EG-Anderson(1) algorithm. Then there exists a positive integer N_0 such that*

$$\min_{N_0+1\leq k\leq N} \|F_\gamma(x_k)\|^2 = O\left(\frac{1}{N}\right).$$

Proof By Theorem 3.1, the sequence $\{x_k\}$ converges to a solution x^* of the problem (3.1.1) and $\lim_{k\to\infty}\|x_k - \tilde{y}_{k+0.5}\| = 0$. Thus, we know that the sequence $\{\tilde{y}_{k+0.5}\}$ also converges to x^*. From the locally Lipschitz continuity of H, there exist an $r \in (0,1)$, an $L^* > 0$ and a positive integer N_0 such that for $k > N_0$, we have $\|x_k - x^*\| \leq r$ and

$$\|H(x_k) - H(\tilde{y}_{k+0.5})\| \leq L^*\|x_k - \tilde{y}_{k+0.5}\|.$$

First, we will discuss the relationship between $\|F_\gamma(x_k)\|$ and $\|F_{t_k}(x_k)\|$ in two cases.

(i) For $t_k = \gamma$, we have

$$\|F_\gamma(x_k)\| = \|F_{t_k}(x_k)\|. \tag{3.1.36}$$

(ii) For $t_k < \gamma$, combining (3.1.29), (3.1.32), (3.1.34) and (3.1.35) yields that

$$\begin{aligned}
\|F_\gamma(x_k)\| &\overset{(3.1.35)}{\leq} \frac{\gamma}{t_k}\|F_{t_k}(x_k)\| \\
&\overset{(3.1.34)}{\leq} \frac{\gamma}{t_k}\|x_k - \tilde{y}_{k+0.5}\| \\
&\overset{(3.1.32)}{<} \gamma\rho^{-1}\mu^{-1}\|H(x_k) - H(\tilde{y}_{k+0.5})\| \\
&\leq \gamma\rho^{-1}\mu^{-1}L^*\|x_k - \tilde{y}_{k+0.5}\| \\
&\overset{(3.1.29)}{\leq} \gamma\rho^{-2}\mu^{-1}L^*\|F_{t_k}(x_k)\|, \quad \forall k > N_0.
\end{aligned} \tag{3.1.37}$$

Combining (3.1.36) and (3.1.37), we know that for any $k > N_0$, we have

$$\|F_\gamma(x_k)\| \leq \max\{\gamma\rho^{-2}\mu^{-1}L^*, 1\}\|F_{t_k}(x_k)\|. \tag{3.1.38}$$

Next, we will prove that $\sum_{k=N_0+1}^{\infty}\|F_{t_k}(x_k)\|^2$ is finite. We only consider the case that both K_{AA} and K_{EG} are infinite, as the analysis is similar or simpler when either of them is finite. For the aforementioned N_0, there exist i_0 and j_0 such that $\{k_i : i \geq i_0\} \cup \{l_j : j \geq j_0\} = \{k : k \geq N_0 + 1\}$. Then we know

$$\sum_{k=N_0+1}^{\infty}\|F_{t_k}(x_k)\|^2 = \sum_{i=i_0}^{\infty}\|F_{t_{k_i}}(x_{k_i})\|^2 + \sum_{j=j_0}^{\infty}\|F_{t_{l_j}}(x_{l_j})\|^2. \tag{3.1.39}$$

From (3.1.22) and (3.1.25), we get

$$0 \leq \sum_{i=i_0}^{\infty}\left(\|x_{k_i} - x^*\|^2 - \|x_{k_i+1} - x^*\|^2\right) + \sum_{i=i_0}^{\infty}\varepsilon_{k_i} \tag{3.1.40}$$

and

$$(1-\mu)\sum_{j=j_0}^{\infty}\|F_{t_{l_j}}(x_{l_j})\|^2 \leq \sum_{j=j_0}^{\infty}\left(\|x_{l_j} - x^*\|^2 - \|x_{l_j+1} - x^*\|^2\right), \tag{3.1.41}$$

respectively.

Adding (3.1.40) and (3.1.41), and using $\|x_{N_0+1} - x^*\| \leq r < 1$, we obtain

$$\begin{aligned}\sum_{j=j_0}^{\infty}\|F_{t_{l_j}}(x_{l_j})\|^2 \leq& \frac{1}{1-\mu}\|x_{N_0+1} - x^*\|^2 + \frac{1}{1-\mu}\sum_{i=0}^{\infty}\varepsilon_{k_i} \\ \leq& \frac{1}{1-\mu} + \frac{1}{1-\mu}M(M+1)\nu^2C_0.\end{aligned} \tag{3.1.42}$$

By (3.1.27), (3.1.39) and (3.1.42), we conclude that

$$\sum_{k=N_0+1}^{\infty}\|F_{t_k}(x_k)\|^2 \leq \tilde{C},$$

where $\tilde{C} := \left(\left(1-\sqrt{\frac{\mu}{2-\mu}}\right)^{-2} + \frac{1}{1-\mu}M(M+1)\right)\nu^2 C_0 + \frac{1}{1-\mu}$.

This together with (3.1.38) yields

$$\sum_{k=N_0+1}^{\infty} \|F_\gamma(x_k)\|^2 \leq \max\{\gamma\rho^{-2}\mu^{-1}L^*, 1\}^2 \sum_{k=N_0+1}^{\infty} \|F_{t_k}(x_k)\|^2 \leq C^*$$

with $C^* := \max\{\gamma\rho^{-2}\mu^{-1}L^*, 1\}^2\tilde{C}$. Hence we conclude that

$$\min_{N_0+1\leq k\leq N} \|F_\gamma(x_k)\|^2 \leq \frac{1}{N-N_0}\sum_{k=N_0+1}^{N} \|F_\gamma(x_k)\|^2 \leq \frac{1}{N-N_0}C^*.$$

□

Remark 3.2. *If the pseudomonotone operator H is L-Lipschitz continuous, we will no longer need the line search step* (3.1.2) *in the EG-Anderson(1) algorithm, in which case we can take t_k to be the constant $t > 0$ satisfying $tL < 1$. In this situation, let $\{x_k\}$ be the sequence generated by the EG-Anderson(1) algorithm, then the following statements hold.*

*(i) (**Sequence convergence**) The sequence $\{x_k\}$ converges to a solution of the problem* (3.1.1)*;*

*(ii) (**Best-iterate convergence rate**)* $\min_{0\leq k\leq N} \|F_t(x_k)\|^2 = O\left(\frac{1}{N}\right)$.

It can be seen that the EG-Anderson(1) algorithm can guarantee the sequence convergence as well as the EG method under the condition $tL < 1$, and we will show that it is faster than the EG method by numerical experiments.

3.2 Subgradient EG-Anderson(1) algorithm and its convergence analysis

The subgradient extra gradient algorithm [15] can be regarded as an improved version of the EG method [44]. It constructs a halfspace at each iteration and replaces the

second projection onto the feasible set with a projection onto this halfspace, which admits a closed-form solution (see Lemma 2.6). Under the same conditions, the subgradient extra gradient algorithm maintains the same convergence properties as the EG method while significantly improving computational efficiency, especially when the feasible set has a complex structure and its projection is expensive to compute.

Based on the above ideas, in this section we present an improved version of the EG-Anderson(1) algorithm, referred to as the subgradient EG-Anderson(1) algorithm. In cases where computing the projection onto the feasible set is challenging, the subgradient EG-Anderson(1) algorithm maintains the convergence behavior of EG-Anderson(1) algorithm and improves efficiency.

3.2.1 Framework of the subgradient EG-Anderson(1) algorithm

The structure of the subgradient EG-Anderson(1) algorithm is as follows:

3.2.2 Convergence analysis

According to the definitions of $y_{k+0.5}$ and y_{k+1} in the subgradient EG-Anderson(1) algorithm, we have

$$y_{k+0.5} - x_k = F_{t_k}(x_k) \quad \text{and} \quad y_{k+1} - x_k = \bar{F}_{t_k}(x_k). \tag{3.2.6}$$

To classify the iteration counts, we partition them into two distinct subsets:

$$K_{GAA} = \{k_0, k_1, \cdots\} \quad \text{and} \quad K_{GEG} = \{l_0, l_1, \cdots\}, \tag{3.2.7}$$

where K_{GAA} consists of iterations determined by (3.2.4), while K_{GEG} contains the remaining iterations defined by (3.2.5).

In the analysis of this section, we assume that the subgradient EG-Anderson(1) algorithm runs infinitely.

Algorithm 2: Subgradient EG-Anderson(1)

Initialization: Choose $x_0 \in \Omega$, $\nu \geq 0, \gamma > 0$, $\tau > \frac{1}{2}$, $\rho, \mu \in (0,1)$ and $\sigma_0 = 1$. Give a sufficiently large $M > 0$.

for $k = 0, 1, 2, \ldots,$ **do**

Step 1: Compute $F_\gamma(x_k) = P_\Omega(x_k - \gamma H(x_k)) - x_k$.

If $F_\gamma(x_k) = \mathbf{0}$, then stop.

Otherwise, let $t_k = \gamma$ and go to **Step 2**.

Step 2: Compute $y_{k+0.5} = P_\Omega(x_k - t_k H(x_k))$ and

$$y_{k+1} = \arg\min_{x \in B_k} \|x - x_k + t_k H(y_{k+0.5})\|^2, \tag{3.2.1}$$

where $B_k = \{x \in \mathbb{R}^n : \langle x_k - t_k H(x_k) - y_{k+0.5}, x - y_{k+0.5} \rangle \leq 0\}$.

If

$$\begin{aligned} & t_k \langle H(y_{k+0.5}) - H(x_k), y_{k+0.5} - y_{k+1} \rangle \\ \leq & \frac{\mu}{2} \left(\|x_k - y_{k+0.5}\|^2 + \|y_{k+0.5} - y_{k+1}\|^2 \right), \end{aligned} \tag{3.2.2}$$

go to **Step 3**.

Otherwise, set $t_k = \rho t_k$ and repeat **Step 2**.

Step 3: Compute $F_{t_k}(x_k) = y_{k+0.5} - x_k$ and $\bar{F}_{t_k}(x_k) = y_{k+1} - x_k$.

If $\|\bar{F}_{t_k}(x_k)\| < \min\{\|F_{t_k}(x_k)\|, \nu \sigma_k^{-\tau}\}$, set

$$\bar{\alpha}_k = \frac{\langle \bar{F}_{t_k}(x_k), \bar{F}_{t_k}(x_k) - F_{t_k}(x_k) \rangle}{\|\bar{F}_{t_k}(x_k) - F_{t_k}(x_k)\|^2}. \tag{3.2.3}$$

Otherwise, set $\bar{\alpha}_k = M + 1$.

Step 4: If $|\bar{\alpha}_k| \leq M$, set

$$x_{k+1} = \bar{\alpha}_k x_k + (1 - \bar{\alpha}_k) y_{k+1}, \quad \sigma_{k+1} = \sigma_k + 1. \tag{3.2.4}$$

Otherwise, set

$$x_{k+1} = y_{k+1}, \quad \sigma_{k+1} = \sigma_k. \tag{3.2.5}$$

end for

Remark 3.3. *In each iteration k, the unique solution y_{k+1} to problem* (3.2.1) *can be equivalently written as*

$$y_{k+1} = P_{B_k}(x_k - t_k H(y_{k+0.5})).$$

Remark 3.4. *Similar to the proof of Lemma 3.2, it can be shown that for each k,*

the line search (3.2.2) *in the subgradient EG-Anderson(1) algorithm terminates after a finite number of iterations.*

Lemma 3.5. *Let* $\{x_k\}$, $\{y_{k+0.5}\}$ *and* $\{y_{k+1}\}$ *be the sequences generated by the subgradient EG-Anderson(1) algorithm and* $x^* \in \mathrm{SOL}(\Omega, H)$. *For every* k, *it holds that*

$$\|y_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - (1-\mu)\|y_{k+0.5} - x_k\|^2 - (1-\mu)\|y_{k+1} - y_{k+0.5}\|^2.$$

Proof Similar to the proof of Lemma 3.4, according to the definition of B_k and the fact that $y_{k+1} \in B_k$, we have

$$\langle x_k - t_k H(x_k) - y_{k+0.5}, y_{k+1} - y_{k+0.5}\rangle \leq 0. \tag{3.2.8}$$

From the definition of $y_{k+0.5}$ in the subgradient EG-Anderson(1) algorithm and Lemma 2.1-(iii), we have

$$\langle x_k - t_k H(x_k) - y_{k+0.5}, x^* - y_{k+0.5}\rangle \leq 0,$$

which implies $x^* \in B_k$. By applying Lemma 2.1-(iii) again, we obtain

$$\langle P_{B_k}(z_k) - z_k, P_{B_k}(z_k) - x^*\rangle \leq 0, \tag{3.2.9}$$

where $z_k = x_k - t_k H(y_{k+0.5})$.

In the first equality of (3.1.17), by utilizing (3.2.8) and following an estimation approach similar to that in (3.1.17), we obtain

$$\langle x_k - t_k H(y_{k+0.5}) - y_{k+0.5}, y_{k+1} - y_{k+0.5}\rangle \leq \frac{\mu}{2}\|x_k - y_{k+0.5}\|^2 + \frac{\mu}{2}\|y_{k+0.5} - y_{k+1}\|^2.$$

The proof can be completed by following the same arguments as in Lemma 3.4, by substituting $y_{k+1} = P_{B_k}(x_k - t_k H(y_{k+0.5}))$ into (3.1.18) and applying (3.2.9) to the second equality in (3.1.18). □

Below, we use Opial's Lemma to prove that the sequence generated by the subgradient EG-Anderson(1) has the global convergence property.

Theorem 3.3. *Let $\{x_k\}$ be the sequence generated by the subgradient EG-Anderson(1) algorithm. Then the sequence $\{x_k\}$ converges to a solution of the problem* (3.1.1).

Proof Following a similar approach to Theorem 3.1, we divide the proof into three steps.

Step 1: $\{x_k\}$ is bounded.

Let x^* be a solution of the problem (3.1.1). If $k_i \in K_{GAA}$, let $\bar{\beta}_{k_i} := 1 - \bar{\alpha}_{k_i}$. Since $\|\bar{F}_{t_{k_i}}(x_{k_i})\| < \|F_{t_{k_i}}(x_{k_i})\|$ when $k_i \in K_{GAA}$, there exists $\tilde{\beta} \in (0, \frac{1}{2}]$ such that

$$\begin{aligned}
&\tilde{\beta}\|\bar{F}_{t_{k_i}}(x_{k_i})\|^2 + (1-2\tilde{\beta})\langle F_{t_{k_i}}(x_{k_i}), \bar{F}_{t_{k_i}}(x_{k_i})\rangle \\
\leq &\tilde{\beta}\|\bar{F}_{t_{k_i}}(x_{k_i})\|^2 + (1-2\tilde{\beta})\|F_{t_{k_i}}(x_{k_i})\|\|\bar{F}_{t_{k_i}}(x_{k_i})\| \\
< &(1-\tilde{\beta})\|F_{t_{k_i}}(x_{k_i})\|^2,
\end{aligned}$$

which implies

$$\langle \bar{F}_{t_{k_i}}(x_{k_i}), \bar{F}_{t_{k_i}}(x_{k_i}) - F_{t_{k_i}}(x_{k_i})\rangle < (1-\tilde{\beta})\|\bar{F}_{t_{k_i}}(x_{k_i}) - F_{t_{k_i}}(x_{k_i})\|^2.$$

By (3.2.3), we have $\bar{\alpha}_{k_i} < 1 - \tilde{\beta}$. Thus $\bar{\beta}_{k_i} = 1 - \bar{\alpha}_{k_i} > \tilde{\beta}$. Using Lemma 3.5, similar to the estimation of (3.1.21), we obtain

$$\begin{aligned}
\|x_{k_i+1} - x^*\|^2 \leq &\|x_{k_i} - x^*\|^2 - (1-\mu)\bar{\beta}_{k_i}\|y_{k_i+0.5} - x_{k_i}\|^2 \\
&- (1-\mu)\bar{\beta}_{k_i}\|y_{k_i+0.5} - y_{k_i+1}\|^2 - \bar{\alpha}_{k_i}\bar{\beta}_{k_i}\|x_{k_i} - y_{k_i+1}\|^2.
\end{aligned} \tag{3.2.10}$$

Combining $|\bar{\alpha}_{k_i}| \leq M$, $\tilde{\beta} < \bar{\beta}_{k_i} < M+1$, $\mu \in (0,1)$ and $\|x_{k_i} - y_{k_i+1}\| = \|\bar{F}_{t_{k_i}}(x_{k_i})\| \leq \nu\sigma_{k_i}^{-\tau} = \nu(\sigma_0 + i)^{-\tau} = \nu(1+i)^{-\tau}$, we conclude that

$$\begin{aligned}
\|x_{k_i+1} - x^*\|^2 &\leq \|x_{k_i} - x^*\|^2 - (1-\mu)\bar{\beta}_{k_i}\|y_{k_i+0.5} - x_{k_i}\|^2 + |\bar{\alpha}_{k_i}||\bar{\beta}_{k_i}|\|x_{k_i} - y_{k_i+1}\|^2 \\
&\leq \|x_{k_i} - x^*\|^2 - (1-\mu)\tilde{\beta}\|y_{k_i+0.5} - x_{k_i}\|^2 + \varepsilon_{k_i} \\
&\leq \|x_{k_i} - x^*\|^2 + \varepsilon_{k_i},
\end{aligned} \tag{3.2.11}$$

where $\varepsilon_{k_i} := M(M+1)\nu^2(1+i)^{-2\tau}$.

If $l_j \in K_{GEG}$, by Lemma 3.5 and $\mu \in (0,1)$, we deduce that

$$\begin{aligned}\|x_{l_j+1} - x^*\|^2 &= \|y_{l_j+1} - x^*\|^2 \\ &\leq \|x_{l_j} - x^*\|^2 - (1-\mu)\|y_{l_j+0.5} - x_{l_j}\|^2 \\ &\quad - (1-\mu)\|y_{l_j+1} - y_{l_j+0.5}\|^2 \\ &\leq \|x_{l_j} - x^*\|^2.\end{aligned} \tag{3.2.12}$$

By setting $\varepsilon_{l_j} = 0$ and combining (3.2.11) and (3.2.12), for every k, we get

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \varepsilon_k, \tag{3.2.13}$$

where $\sum_{k=0}^{\infty} \varepsilon_k = \sum_{i=0}^{\infty} \varepsilon_{k_i} < \infty$ holds due to $\tau > \frac{1}{2}$. From (3.2.13), we have

$$\|x_k - x^*\|^2 \leq \|x_0 - x^*\|^2 + \sum_{k=0}^{\infty} \varepsilon_k < \infty.$$

Hence $\{x_k\}$ is bounded.

Step 2: any cluster point of $\{x_k\}$ belongs to $\mathrm{SOL}(\Omega, H)$, i.e. a solution of the problem (3.1.1).

Due to the boundedness of $\{x_k\}$, it has at least one cluster point $\tilde{x}$. Let $\{x_{q_j}\}$ be a convergent subsequence such that $\lim_{j\to\infty} x_{q_j} = \tilde{x}$. We now verify $\tilde{x} \in \mathrm{SOL}(\Omega, H)$ in two steps:

Step 2.1: $\lim_{k\to\infty} \|F_{t_k}(x_k)\| = 0$.

Similar to the proof in Theorem 3.1, we only focus on the case where both K_{GAA} and K_{GEG} are infinite, as the situation where one of them is finite would be simpler.

Rearranging the terms in (3.2.11) and (3.2.12) respectively, we infer that

$$(1-\eta)\tilde{\beta}\|y_{k_i+0.5} - x_{k_i}\|^2 \leq \|x_{k_i} - x^*\|^2 - \|x_{k_i+1} - x^*\|^2 + \varepsilon_{k_i} \tag{3.2.14}$$

and

$$0 \leq \|x_{l_j} - x^*\|^2 - \|x_{l_j+1} - x^*\|^2. \tag{3.2.15}$$

Combining (3.2.14) and (3.2.15) with $\mu \in (0,1)$ and $\tilde{\beta} \in (0, \frac{1}{2}]$, we have

$$\begin{aligned}\sum_{i=0}^{\infty} \|F_{t_{k_i}}(x_{k_i})\|^2 =& \sum_{i=0}^{\infty} \|y_{k_i+0.5} - x_{k_i}\|^2 \\ \leq& \frac{1}{(1-\mu)\tilde{\beta}}\|x_0 - x^*\|^2 + \frac{1}{(1-\mu)\tilde{\beta}} \sum_{i=0}^{\infty} \varepsilon_{k_i} < \infty.\end{aligned} \tag{3.2.16}$$

Using (3.2.11) and (3.2.12), a similar estimation for (3.2.16) can yield

$$\sum_{j=0}^{\infty} \|F_{t_{l_j}}(x_{l_j})\|^2 \leq \frac{1}{1-\mu}\|x_0 - x^*\|^2 + \frac{1}{1-\mu} \sum_{i=0}^{\infty} \varepsilon_{k_i} < \infty. \tag{3.2.17}$$

Combining (3.2.16) and (3.2.17), we deduce that

$$\sum_{k=0}^{\infty} \|F_{t_{k_i}}(x_{k_i})\|^2 = \sum_{i=0}^{\infty} \|F_{t_{k_i}}(x_{k_i})\|^2 + \sum_{j=0}^{\infty} \|F_{t_{l_j}}(x_{l_j})\|^2 < \infty,$$

which implies $\lim_{k\to\infty} \|F_{t_k}(x_k)\| = 0$.

Step 2.2: $\lim_{k\to\infty} \|F_\gamma(x_k)\| = 0$.

Let $\tilde{y}_{k+0.5} := P_\Omega(x_k - t_k\rho^{-1}H(x_k))$ and $\tilde{y}_{k+1} := P_{B_k}(x_k - t_k\rho^{-1}H(\tilde{y}_{k+0.5}))$. Following the same proof method as in Step 2.2 of Theorem 3.1, we can obtain $\lim_{k\to\infty} \|F_\gamma(x_k)\| = 0$. Furthermore, combining this with Lemma 3.1, we conclude that $\tilde{x} \in \mathrm{SOL}(\Omega, H)$.

Step 3: $\{x_k\}$ is convergent to a solution of the problem (3.1.1).

From (3.2.13) and Lemma 2.5, it follows that $\lim_{k\to\infty} \|x_k - x^*\|^2$ exists for any $x^* \in \mathrm{SOL}(\Omega, H)$. Combining with the result in Step 2 that any cluster point of $\{x_k\}$ is a solution of the problem (3.1.1), we conclude the proof by applying Lemma 2.3. □

If H is locally Lipschitz continuous at any solution of the problem (3.1.1), the following conclusion regarding the convergence rate of the residual function still holds.

Theorem 3.4. *(Best-iterate convergence rate) Suppose that H is locally Lipschitz continuous at any solution of the problem* (3.1.1). *Let $\{x_k\}$ be the sequence generated by the subgradient EG-Anderson(1) algorithm. Then there exists a positive integer $\bar{N}_0$ such that*

$$\min_{\bar{N}_0+1\leq k\leq N} \|F_\gamma(x_k)\|^2 = O\left(\frac{1}{N}\right).$$

Proof Using the similar estimation approach as in (3.1.38), we deduce that there exists a positive integer $\bar{N}_0$ such that for all $k > \bar{N}_0$,

$$\|F_\gamma(x_k)\| \leq \max\{\gamma\rho^{-2}\eta^{-1}L^*, 1\}\|F_{t_k}(x_k)\|,$$

where L^* is the local Lipschitz constant of H at the solutions of the problem (3.1.1).

We only consider the case where both K_{GAA} and K_{GEG} are infinite. For the aforementioned $\bar{N}_0$, there exist $\bar{I}_0$ and $\bar{J}_0$ such that

$$\{k_i : i \geq \bar{I}_0\} \cup \{l_j : j \geq \bar{J}_0\} = \{k : k \geq \bar{N}_0 + 1\}.$$

Using (3.2.11) and (3.2.12), and following the similar estimation approach as in (3.2.16) and (3.2.17), we obtain

$$\sum_{k=\bar{N}_0+1}^{\infty} \|F_{t_k}(x_k)\|^2 = \sum_{i=\bar{I}_0}^{\infty} \|F_{t_{k_i}}(x_{k_i})\|^2 + \sum_{j=\bar{J}_0}^{\infty} \|F_{t_{l_j}}(x_{l_j})\|^2 < \infty.$$

The proof can be completed by the same way in Theorem 3.2. □

3.3 Numerical experiments

In this section, we perform some numerical examples to compare the EG-Anderson(1) with Anderson(1) [7] and the EG method [44]. All the codes were written in Matlab (R2023b) and run on a MacBook Air (16.00GB of RAM).

In the following numerical experiments, the stopping rule for Examples 3.1-3.4 is set by

$$r(x_k) := \|F_1(x_k)\| = \|x_k - P_\Omega(x_k - H(x_k))\| < 10^{-8}$$

or the maximum iteration exceeds 10^4 times. The parameters in the EG-Anderson(1) are set as follows

$$\nu = 30, M = 5000, \tau = 0.6, \rho = 0.8, \mu = 0.5.$$

In the figures and tables of this section, 'Sec.' represents the CPU time in seconds and 'Iter.' represents the number of iterations. Specifically, the numbers in parentheses represent the number of Anderson step (3.1.4) executed. Moreover, '\' indicates that the number of iterations exceeds 10^4, and the corresponding CPU time is not counted, represented by '–'. Furthermore, the best performing algorithm in terms of the average number of iterations and CPU time is highlighted in bold for each combination of dimension n and parameter γ.

Example 3.1. *[30] Consider the Harker-Pang problem with linear mapping* $H(x) := Wx + w_0$, *where* $w_0 \in \mathbb{R}^n$ *and*

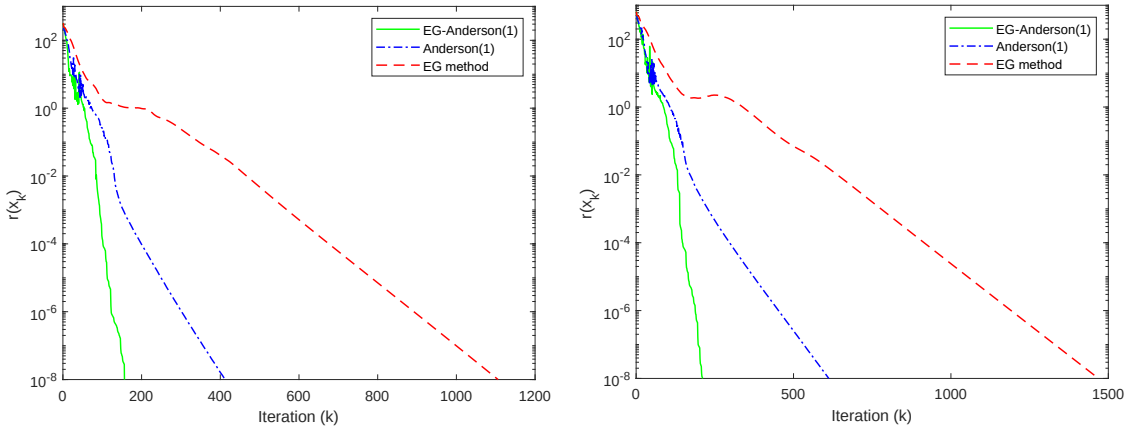
$$W := A^{\mathrm{T}}A + S + D.$$

Here, A *is an* $n \times n$ *matrix,* S *is an* $n \times n$ *skew-symmetric matrix and* D *is an* $n \times n$ *diagonal matrix with nonnegative diagonal entries. Therefore, it follows that* W *is positive semidefinite. Let the feasible set be* $\Omega := \{x \in \mathbb{R}^n : \mathbf{0} \leq x \leq 20\mathbf{e}\}$. *It is clear that* H *is monotone and Lipschitz continuous.*

We can easily obtain that the Lipschitz constant of H is $L = \|W\|$. Applying the EG-Anderson(1) to this example, instead of using line search, we can do experiment with a constant stepsize t that satisfies $tL < 1$.

In the following experiments, we let $t = \frac{0.7}{L}$, and every entry of the skew-symmetric matrix S is uniformly generated from $(-5, 5)$, and every diagonal entry of D is uniformly generated from $(0, 2)$, and A, w_0 are randomly generated.

Figure 3.1 compares the decreasing on the residual function by the EG-Anderson(1), Anderson(1) and the EG algorithms at the same random initial point for Example 3.1 with $n = 500$ and $n = 2000$, respectively. For different dimension n, Table 3.1 illustrates the average number of iterations and CPU time of the corresponding experiments at ten random initial points, where we see that the superiorities of the EG-Anderson(1) over Anderson(1) and the EG algorithms gradually emerges as the dimension increases.



(a) $n = 500$ (b) $n = 2000$

Figure 3.1: Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.1

Example 3.2. *[62] Consider the quadratic fractional programming problem*

$$\begin{aligned} \min \quad & \varphi(x) := \frac{x^T Q x + a^T x + a_0}{b^T x + b_0}, \\ \text{s.t.} \quad & x \in \Omega := \{x \in \mathbb{R}^n : 2\mathbf{e} \leq x \leq 10\mathbf{e}\} \end{aligned}$$

Sec.(avr) Iter.(avr)	EG-Anderson(1)	Anderson(1)	EG
$n = 100$	0.0038 **97.8** (96.6)	**0.0028** 226.4	0.0040 501.7
$n = 500$	**0.0233** **169** (166.4)	0.0243 402.6	0.1171 1103.8
$n = 1000$	**0.0972** **197.5** (195.6)	0.1172 545.8	0.5764 1417.8
$n = 2000$	**0.3462** **214.5** (211.6)	0.4789 595.4	2.2768 1465.2
$n = 5000$	**2.2158** **244** (240.6)	2.7644 600.9	14.1645 1538
$n = 10000$	**9.6154** **260.3** (256.7)	10.5268 563.4	63.4244 1685

Table 3.1: Comparisons of the three algorithms for Example 3.1

with

$$Q := Q_0^{\mathrm{T}} Q_0 + I, a := \mathbf{e} + c, b := \mathbf{e} + d, a_0 := 1 + c_0, b_0 := 1 + d_0,$$

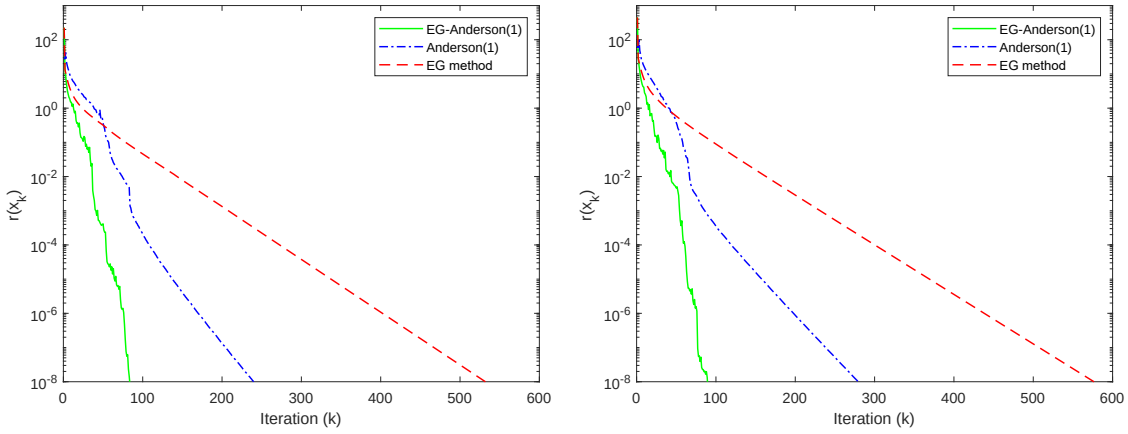
where $I \in \mathbb{R}^{n\times n}$ *is the identity matrix, and* $Q_0 \in \mathbb{R}^{n\times n}$, $c, d \in \mathbb{R}^n$, $c_0, d_0 \in \mathbb{R}$ *are randomly generated from a uniform distribution.*

It is easily verified that $\Omega \subseteq \{x \in \mathbb{R}^n : b^T x + b_0 > 0\}$ *and* Q *is positive definite, and consequently* φ *is pseudoconvex on* Ω. *Thus,* $H(x) := \nabla\varphi(x)$ *in the problem* (3.1.1) *can be written in the following explicit form:*

$$H(x) = \frac{(b^T x + b_0)(2Qx + a) - b(x^T Qx + a^T x + a_0)}{(b^T x + b_0)^2}.$$

Let $\gamma = 0.6$. Starting from the same random initial point, Figure 3.2 shows the comparisons of the results obtained by the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.2 with $n = 500$ and $n = 2000$. Table 3.2 presents the average number of iterations and CPU time for the experiments with more cases on

n, conducted with ten different random initial points. Thus, the results indicate that the EG-Anderson(1) outperforms both Anderson(1) and the EG algorithms across various dimensions in terms of iterations and CPU time.



(a) $n = 500$ (b) $n = 2000$

Figure 3.2: Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.2

Example 3.3. *[9] Consider the following nonlinear complementarity problem (NCP)*

$$H(x) \geq \mathbf{0}, \quad x \geq \mathbf{0}, \quad x^{\mathrm{T}} H(x) = 0, \tag{3.3.1}$$

where $H : \mathbb{R}^n \to \mathbb{R}^n$ *is defined by*

$$H(x) = (e^{-x^T U x} + \kappa)(Px + \iota).$$

Here U *is an* $n \times n$ *positive definite matrix,* P *is an* $n \times n$ *positive semidefinite matrix,* $\iota \in \mathbb{R}^n$ *and* $\kappa > 0$*. It is also shown that* H *is pseudomonotone [9]. Furthermore, the NCP in* (3.3.1) *can be equivalently formulated as the problem* (3.1.1) *with* $\Omega := \{x \in \mathbb{R}^n : x \geq \mathbf{0}\}$.

In numerical tests, we take $\kappa = 0.01, P = P_0^T P_0, U = U_0^T U_0$, where matrices $P_0, U_0 \in \mathbb{R}^{n \times n}$ and vector $\iota \in \mathbb{R}^n$ are randomly generated as follows with an input

Sec.(avr) Iter.(avr)	EG-Anderson(1)	Anderson(1)	EG
$n = 100$	0.0051 **99** (99)	**0.0044** 299.7	0.0105 609
$n = 500$	**0.0439** **93.2** (91.4)	0.0446 238.5	0.1819 531
$n = 1000$	**0.1631** **102.6** (101.3)	0.1937 285.2	0.7851 588.8
$n = 2000$	**0.6839** **101.2** (99.9)	0.8337 285.3	3.4400 576
$n = 5000$	**5.5632** **95.3** (93.7)	6.9468 276	27.9809 538
$n = 10000$	**18.7886** **102.3** (100.9)	20.7397 262	89.8027 556

Table 3.2: Comparisons of the three algorithms for Example 3.2

integer n:

$$P_0 = \text{randn}(n, n); U_0 = \text{randn}(n, n); \hat{x} = \max\left(\mathbf{0}, \text{randn}(n, 1)\right);$$

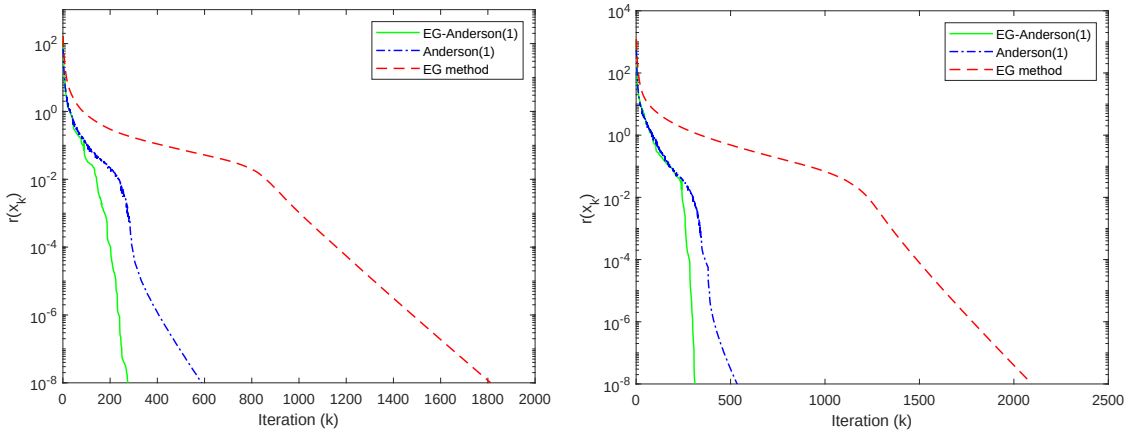
$$\iota = (-P * \hat{x}). * (\hat{x} > 0) + (-P * \hat{x} + \text{rand}(n, 1)). * (\hat{x} == 0).$$

From the above generation, we know that $\hat{x}$ is a solution of (3.3.1).

In Figure 3.3, we present a comparative analysis of the outcomes achieved by applying the EG-Anderson(1), Anderson(1) and the EG algorithms to Example 3.3. All three methods in Figure 3.3 are initialized with the same random initial point. Table 3.3 summarizes the average number of iterations and CPU time for each method, where the corresponding experiments are repeated ten times with different random initial points.

In all dimensions, the average number of iterations and CPU time of the EG-Anderson(1) are significantly less than those of Anderson(1) and the EG algorithms. The advantages of the EG-Anderson(1) become more pronounced as the dimension

increases. For instance, in the case of $n = 10000$, the EG-Anderson(1) achieves an average number of iterations of 552.2 and an average CPU time of 31.5962 seconds, significantly outperforming Anderson(1) and the EG algorithms.



(a) $n = 500$ (b) $n = 2000$

Figure 3.3: Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.3

Example 3.4. *Consider the following partial differential equation (PDE) problem with free boundary*

$$\begin{aligned} -\triangle u + \frac{9}{(1-p)^2}u^p + \delta e^{-u} &= c(\xi,\varsigma) \quad \text{in } \Lambda_+, \\ u &= 0 \qquad\qquad\quad \text{in } \Lambda_0, \\ u = |\nabla u| &= 0 \qquad\qquad\quad \text{on } \Gamma, \\ u &= v(\xi,\varsigma) \qquad\quad \text{on } \partial\Lambda, \end{aligned} \tag{3.3.2}$$

where $p \in (0,1)$, $\delta \geq 0$, $\Lambda = (0,1)\times(0,1)$, $\Lambda_+ = \{(\xi,\varsigma) \in \Lambda : u(\xi,\varsigma) > 0\}$, $\Lambda_0 = \{(\xi,\varsigma) \in \Lambda : u(\xi,\varsigma) = 0\}$ *and* $\Gamma = \partial\Lambda_0 = \partial\Lambda_+ \cap \Lambda$ *are unknown. Let* $r^2 = \xi^2 + \varsigma^2$. *We choose*

$$c(\xi,\varsigma) = D(r,p)R(r,p) + \delta e^{-R(r,p)}$$

Sec.(avr) Iter.(avr)	EG-Anderson(1)	Anderson(1)	EG
$n = 100$	0.0041	**0.0027**	0.0090
$\gamma = 0.27$	**128.2** (125.4)	256.7	710.3
$n = 500$	**0.0405**	0.0424	0.2523
$\gamma = 0.03$	**253.2** (249)	566.2	1754.4
$n = 1000$	**0.1802**	0.1823	1.0553
$\gamma = 0.02$	**232.9** (228.8)	504	1513.1
$n = 2000$	**0.9466**	0.9478	6.1649
$\gamma = 0.008$	**321.2** (316.6)	622.3	2122.6
$n = 5000$	**7.3880**	8.3522	63.3128
$\gamma = 0.002$	**485.2** (478.6)	1097	4146.3
$n = 10000$	**31.5962**	33.8899	287.9682
$\gamma = 0.0009$	**552.2** (544.5)	1173.7	4836.6

Table 3.3: Comparisons of the three algorithms for Example 3.3

and

$$v(\xi, \varsigma) = R(r, p),$$

where

$$D(r,p) := -\frac{3(3-p)\left[(3r-1)(1-p)+6r\right]}{r(1-p)^2(3r-1)^2} + \frac{27}{2(1-p)^2}\left(\frac{2}{3}\right)^p\left(\frac{3r-1}{2}\right)^{p-3}$$

and $R(r,p) := \left(\frac{3r-1}{2}\right)^{\frac{2}{1-p}} \max\left(0, r-\frac{1}{3}\right).$

Then problem (3.3.2) has a solution as follows

$$u(\xi, \varsigma) = R(r,p) = \left(\frac{3r-1}{2}\right)^{\frac{2}{1-p}} \max\left(0, r-\frac{1}{3}\right).$$

Dividing the interval (0,1) into N subintervals of equal width h provides mesh points (ξ_i, ς_j) where

$$\xi_i = ih, \quad i = 0, 1, \ldots, N,$$

$$\varsigma_j = jh, \quad j = 0, 1, \ldots, N.$$

Using the five point finite difference method for the problem (3.3.2) at grid (ξ_i, ς_j) gives

$$-u_{i,j+1} - u_{i,j-1} + 4u_{i,j} - u_{i+1,j} - u_{i-1,j} + \frac{9h^2}{(1-p)^2} u_{i,j}^p + h^2 \delta e^{-u_{i,j}} = h^2 c_{i,j}, \quad (\xi_i, \varsigma_j) \in \Lambda_+$$

and

$$u_{i,j} = v_{i,j}, \qquad (\xi_i, \varsigma_j) \in \partial\Lambda.$$

Let $x := (u_{1,1}, u_{2,1}, \ldots, u_{N-1,1}, u_{1,2}, \ldots, u_{N-1,N-1})^{\mathrm{T}} \in \mathbb{R}^{(N-1)^2}$ and $\tilde{c}, \tilde{v} \in \mathbb{R}^{(N-1)^2}$ be the corresponding vectors transformed by $c_{i,j}$ and $v_{i,j}$. Then, we obtain an NCP with

$$H(x) := Bx + E \max(\mathbf{0}, x)^p + Vf(x) + q,$$

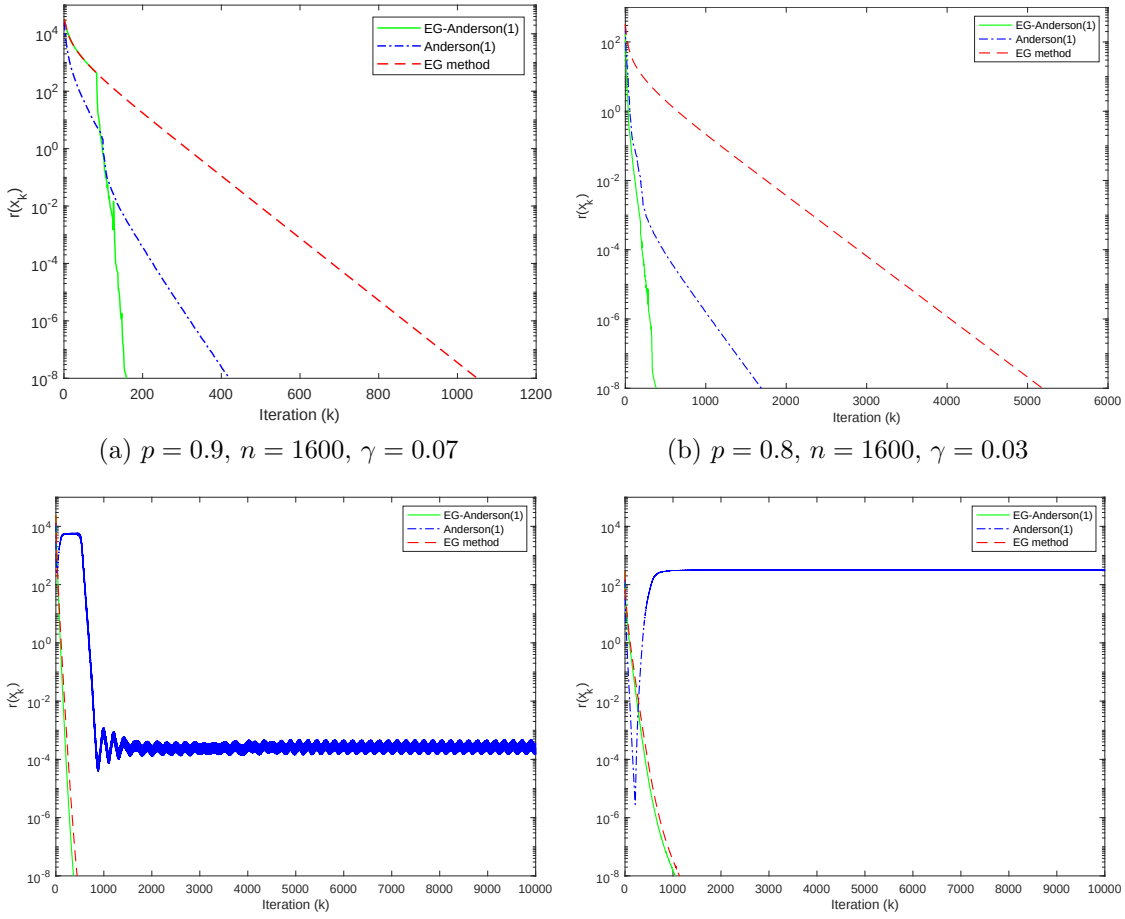
where B is a block tri-diagonal positive definite matrix of dimension $(N-1)^2 \times (N-1)^2$, E and V are both $(N-1)^2 \times (N-1)^2$ dimensional diagonal matrices with the diagonal elements being $\frac{9h^2}{(1-p)^2}$ and δh^2 respectively, $f(x) := (e^{-u_{1,1}}, e^{-u_{2,1}}, \ldots, e^{-u_{N-1,N-1}})^{\mathrm{T}}$ $\in \mathbb{R}^{(N-1)^2}$ and $q = -h^2\tilde{c} - \tilde{v} \in \mathbb{R}^{(N-1)^2}$. Note that the dimension of the corresponding complementarity problem is $n = (N-1)^2$. Furthermore, the NCP can be equivalently formulated as the problem (3.1.1) with $\Omega := \{x \in \mathbb{R}^n : x \geq \mathbf{0}\}$.

In the experiments, let $\delta = 1$. We set $p = 0.9$ and $p = 0.8$, respectively. The effectiveness of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.4 with $n = 900$ and $n = 1600$ is compared in Figure 3.4. All three methods are started from the same random initial point for each case.

As can be seen from Table 3.4, the average number of iterations for the EG-Anderson(1) is consistently smaller than that of Anderson(1) and the EG algorithms in all dimensions. Additionally, the EG-Anderson(1) has a shorter CPU time, highlighting its higher computational efficiency. In high-dimensional problems, such as $n = 8100$ and $n = 10000$, the EG-Anderson(1) outperforms the other two algorithms

in terms of the average number of iterations and CPU time, demonstrating its higher efficiency and performance.

Table 3.5 presents the average number of iterations and CPU time required for each algorithm in Example 3.4 with a different value of $p = 0.8$.



(a) $p = 0.9$, $n = 1600$, $\gamma = 0.07$

(b) $p = 0.8$, $n = 1600$, $\gamma = 0.03$

(c) $p = 0.9, n = 900, \gamma = 0.39$

(d) $p = 0.8, n = 900, \gamma = 0.39$

Figure 3.4: Comparisons of the convergence behaviours of the EG-Anderson(1), Anderson(1) and the EG algorithms for Example 3.4

Example 3.5. *Consider the following linear complementarity problem (LCP)*

$$\tilde{M}x + \tilde{q} \geq \mathbf{0}, \quad x \geq \mathbf{0}, \quad x^{\mathrm{T}}(\tilde{M}x + \tilde{q}) = 0, \tag{3.3.3}$$

Sec.(avr) Iter.(avr)	EG-Anderson(1)	Anderson(1)	EG
$n = 100$	**0.0036**	0.0056	0.0135
$\gamma = 0.01$	**64.7** (47.7)	280	530
$n = 900$	0.0165	**0.0158**	0.0604
$\gamma = 0.06$	**121.1** (73.1)	300.3	656
$n = 1600$	**0.0236**	0.0246	0.1022
$\gamma = 0.07$	**147.2** (79.2)	398.6	964
$n = 2500$	**0.0349**	0.0412	0.1740
$\gamma = 0.08$	**186.9** (97.9)	526.7	1288
$n = 4900$	**0.0857**	0.1050	0.5587
$\gamma = 0.09$	**296.8** (156.8)	771.8	2178
$n = 8100$	**0.1693**	0.1972	1.0951
$\gamma = 0.1$	**421.3** (228.3)	1064.9	3181
$n = 10000$	**0.2659**	0.3225	1.8438
$\gamma = 0.1$	**502.9** (276.9)	1270.7	3899

Table 3.4: Comparisons of the three algorithms for Example 3.4 with $p = 0.9$

where $\tilde{M}$ *is an* $n \times n$ *P-matrix and* $\tilde{q} \in \mathbb{R}^n$ *is a vector. Additionally, the LCP in* (3.3.3) *can be equivalently expressed as a nonmonotone* $\mathrm{VI}(\Omega, H)$*, where* $H(x) = \tilde{M}x + \tilde{q}$ *and* $\Omega = \{x \in \mathbb{R}^n : x \geq \mathbf{0}\}$.

Note that an $n \times n$ matrix $\tilde{M} = (m_{ij})$ is called a P-matrix if all principal minors of $\tilde{M}$ are positive. We refer to Example 4.4 from [17] to generate the matrix $\tilde{M}$ and vector $\tilde{q}$. First, we randomly generate a dense matrix $\tilde{A} \in \mathbb{R}^{n\times n}$ and a vector $\tilde{q} \in \mathbb{R}^n$, with elements uniformly distributed in the range $(-5, 5)$. By applying the QR decomposition to $\tilde{A}$, we obtain an upper triangular matrix $\tilde{N}$. Next, we replace the diagonal elements of $\tilde{N}$ with their absolute values, resulting in a triangular matrix $\tilde{M}$ with positive diagonal entries. This ensures that $\tilde{M}$ is a P-matrix.

Below, we compare the performance of the EG-Anderson(1) algorithm, iPCA

Sec.(avr) Iter.(avr)	EG-Anderson(1)	Anderson(1)	EG
$n = 100$	0.0115	**0.0088**	0.0426
$\gamma = 0.005$	**371.2** (371.2)	774	2448.3
$n = 900$	**0.0383**	0.0683	0.3788
$\gamma = 0.02$	**376.4** (376.4)	1478.4	4512.3
$n = 1600$	**0.0843**	0.1161	0.8202
$\gamma = 0.03$	**375.1** (375.1)	1427	5183.8
$n = 2500$	**0.0640**	0.1019	0.6231
$\gamma = 0.05$	**397.9** (397.9)	1417.5	4761.1
$n = 4900$	**0.1355**	0.2959	1.8412
$\gamma = 0.06$	**474.4** (474.4)	2275.9	7554.5
$n = 8100$	**0.2472**	0.4831	–
$\gamma = 0.07$	**639.4** (638.4)	2692.5	\
$n = 10000$	**0.3341**	0.6177	–
$\gamma = 0.08$	**671.7** (669.7)	2660.8	\

Table 3.5: Comparisons of the three algorithms for Example 3.4 with $p = 0.8$

from [23], IPA$_L$ from [76] and the semi-smooth Newton algorithm[1] from [26] for solving (3.3.3). The stopping criterion for the experiments is the same as that for the semi-smooth Newton algorithm, specifically,

$$0.5\|s_k - x_k - (\tilde{M}x_k + \tilde{q})\|^2 < 10^{-8}$$

or the maximum iteration exceeds 10^4 times, where $s_k = ((s_k)_1, (s_k)_2, \dots)^{\mathrm{T}}$ with $(s_k)_i = \sqrt{(x_k)_i^2 + (\tilde{M}x_k + \tilde{q})_i^2}$, $i = 1, \dots, n$. Note that for an n-dimensional vector z, $(z)_i$ represents the i-th component of z.

We can find that the Lipschitz constant of H is $\|\tilde{M}\|$. When we run the EG-Anderson(1) to this example, rather than employing a line search, we do experiment with a constant step size t that meets the condition $t\|\tilde{M}\| < 1$. In the following

[1] The solver, developed by Y. Tassa, https://www.mathworks.com/matlabcentral/fileexchange/20952-lcp-mcp-solver-newton-based

experiments, we let $t = \frac{0.7}{\|\tilde{M}\|}$. In both the iPCA and the IPA$_L$, the step size is taken to be the same as that in the EG-Anderson(1) algorithm. In addition, we set the parameter $\gamma = 1.5$ in the iPCA.

Table 3.6 presents the average number of iterations and CPU time required by the four algorithms to meet the stopping criterion for (3.3.3) across different dimensions, starting from ten random initial points. It can be observed that the EG-Anderson(1) algorithm consistently outperforms the other three algorithms in terms of CPU time across various dimensions.

Sec.(avr) Iter.(avr)	EG-Anderson(1)	semi-smooth Newton	iPCA	IPA$_L$
$n = 100$	**0.0038** 64 (63.2)	0.0237 **29.4**	0.0044 149	0.0215 281.5
$n = 500$	**0.0188** 125.2 (124.8)	0.3223 **45.7**	0.0464 286.8	0.4225 638.7
$n = 1000$	**0.1372** 293.6 (288.2)	1.3163 **50.8**	0.7139 1221.2	19.1573 2718.1
$n = 2000$	**0.5563** 306.4 (293.8)	8.0195 **68.9**	2.3217 913.1	32.1018 2048.6
$n = 5000$	**4.3059** **408.3** (337.7)	17.1193 578.4	959.8955 976.8	854.2569 4994.4

Table 3.6: Comparisons of the four algorithms for Example 3.5

Chapter 4

A Stochastic Accelerated Algorithm for Pseudomonotone Variational Inequalities via Anderson Acceleration

In this chapter, we propose a stochastic accelerated algorithm by utilizing the Anderson acceleration method to solve pseudomonotone SVI $(\Omega, \mathbb{E}[T(\xi, \cdot)])$, which does not require prior knowledge of the Lipschitz constant L of $\mathbb{E}[T(\xi, \cdot)]$ nor the contractivity of the fixed point mapping.

We begin by reviewing the SVIs. By combining the EG method with Anderson acceleration and the SA method, we present the framework of the proposed stochastic algorithm in section 4.1 and introduce some definitions and properties used in the subsequent convergence analysis. Then, in this section, we also analyze the convergence properties of the proposed stochastic algorithm, providing almost sure convergence of the sequence, the convergence rate of the mean residual function, and the oracle complexity. Finally, in section 4.2, we conduct numerical experiments on two examples to compare the numerical performance of the proposed stochastic algorithm with that of the stochastic EG method. Additionally, we perform numerical experiments on portfolio selection problems with real data, comparing the perfor-

mance of the portfolio solving by the proposed stochastic algorithm with that of the naive evenly weighted portfolio.

4.1 Stochastic EG-Anderson(1) algorithm and its convergence analysis

We begin with a review of SVI $(\Omega, \mathbb{E}[T(\xi, \cdot)])$. Let $(\Xi, \mathcal{G})$ be a measurable space and $T : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ a measurable (random) operator. Given a nonempty closed convex set Ω, the SVI is to find an $x^* \in \Omega$ such that

$$\langle \mathbb{E}[T(\xi, x^*)], x - x^* \rangle \geq 0, \quad \forall x \in \Omega, \tag{4.1.1}$$

where $\xi : X \to \Xi$ is a random variable defined on a probability space $(X, \mathcal{F}, \mathbb{P})$. In this thesis, we refer to the solution set of the problem (4.1.1) as $\mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$, and we assume that $\mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)]) \neq \emptyset$.

For any $x \in \mathbb{R}^n$ and $t > 0$, we denote the residual function by

$$r_t(x) := \left\| P_\Omega \left(x - t\mathbb{E}[T(\xi, x)] \right) - x \right\|. \tag{4.1.2}$$

The following lemma demonstrates that the residual function (4.1.2) can be employed to quantify solutions to $\mathrm{SVI}(\Omega, \mathbb{E}[T(\xi, \cdot)])$.

Lemma 4.1. [25, Proposition 1.5.8] *$x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$ if and only if it is a solution of $r_t(x) = 0$, where t can be any positive number.*

4.1.1 Framework of the stochastic EG-Anderson(1) algorithm

Before presenting the framework of the proposed stochastic algorithm, we first introduce some key notations.

We define the oracle error $\epsilon : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$\epsilon(\xi, x) := T(\xi, x) - \mathbb{E}[T(\xi, x)], \quad \forall \xi \in \Xi, \quad \forall x \in \mathbb{R}^n. \tag{4.1.3}$$

Given $N \in \mathbb{N}$, let $\xi^N = \{\xi_j\}_{j=1}^N$ be an i.i.d. sample from Ξ. We define the empirical mean operator and the oracle's empirical mean error associated to ξ^N by

$$\hat{H}(\xi^N, x) := \frac{1}{N}\sum_{j=1}^{N} T(\xi_j, x) \quad \text{and} \quad \hat{\epsilon}(\xi^N, x) := \frac{1}{N}\sum_{j=1}^{N} \epsilon(\xi_j, x) \tag{4.1.4}$$

respectively. In this section, we propose a stochastic algorithm based on the EG-Anderson(1) algorithm to solve the problem (4.1.1). The following Algorithm 3 presents the proposed algorithm, called the stochastic EG-Anderson(1) algorithm, in which the line search framework from [74] is utilized in Step 2.

We will analyze the convergence properties of the stochastic EG-Anderson(1) algorithm under the following assumptions.

Assumption 4.1. *The mean operator $\mathbb{E}[T(\xi, \cdot)] : \mathbb{R}^n \to \mathbb{R}^n$ is pseudomonotone on Ω.*

Remark 4.1. *[41] Suppose $T(\xi, \cdot)$ is pseudomonotone on Ω in an almost sure sense. Then $\mathbb{E}[T(\xi, \cdot)]$ is pseudomonotone on Ω. However, the reverse implication does not hold in general. For example, consider $T(\xi, x) = \begin{cases} x, & \text{if } \xi = 0, \\ -x, & \text{if } \xi = 1. \end{cases}$ It is easy to see that $\mathbb{E}[T(\xi, x)] = \mathbf{0}$ is pseudomonotone, while $T(\xi, x) = -x$ is not pseudomonotone for $\xi = 1$.*

Assumption 4.2. *(i) There exists a measurable function $L : \Xi \to \mathbb{R}$ such that $L(\xi) \geq 1$ holds with probability 1, and for any $x, y \in \mathbb{R}^n$,*

$$\|T(\xi, x) - T(\xi, y)\| \leq L(\xi)\|x - y\|.$$

(ii) There exist $\check{x}$ and $p \geq 2$ such that $\mathbb{E}[\|T(\xi, \check{x})\|^{2p}] < \infty$ and $\mathbb{E}[L(\xi)^{2p}] < \infty$.

Assumption 4.3. *Let $\xi^k := \{\xi_j^k : j \in [S_k]\}$ and $\eta^k := \{\eta_j^k : j \in [S_k]\}$. For all k, $\{\xi^k\}$ and $\{\eta^k\}$ are i.i.d. samples of ξ such that $\{\xi^k\}$ and $\{\eta^k\}$ are independent of each other. Moreover, $\sum_{k=0}^{\infty} \frac{1}{\sqrt{S_k}} < \infty$.*

Algorithm 3: Stochastic EG-Anderson(1)

Initialization: Choose $x_0 \in \Omega$, $\nu \geq 0, \gamma \in (0,1]$, $\rho \in (0,1)$, $\mu \in (0, \frac{\sqrt{3}}{3})$, $\tau > \frac{1}{2}$, $\sigma_0 = 1$ and the sample rate $\{S_k\}$. Give a sufficiently large $M > 0$.

for $k = 0, 1, 2, \ldots,$ **do**

Step 1: Let $t_k = \gamma$ and generate samples $\xi^k := \{\xi_j^k\}_{j=1}^{S_k}$ of the random variable ξ. Compute $\hat{H}(\xi^k, x_k) = \frac{1}{S_k}\sum_{j=1}^{S_k} T(\xi_j^k, x_k)$ and

$$F_\gamma(\xi^k, x_k) = P_\Omega\left(x_k - \gamma \hat{H}(\xi^k, x_k)\right) - x_k.$$

If $F_\gamma(\xi^k, x_k) = \mathbf{0}$, set

$$x_{k+1} = x_k, \quad \sigma_{k+1} = \sigma_k. \tag{4.1.5}$$

Otherwise, go to **Step 2**.

Step 2: Compute $y_{k+0.5} = P_\Omega\left(x_k - t_k \hat{H}(\xi^k, x_k)\right)$ and $\hat{H}(\xi^k, y_{k+0.5}) = \frac{1}{S_k}\sum_{j=1}^{S_k} T(\xi_j^k, y_{k+0.5})$.

If

$$t_k \|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5})\| \leq \mu \|x_k - y_{k+0.5}\|, \tag{4.1.6}$$

go to **Step 3**.

Otherwise, set $t_k = \rho t_k$ and repeat **Step 2**.

Step 3: Compute $F_{t_k}(\xi^k, x_k) = y_{k+0.5} - x_k$. Generate samples $\eta^k := \{\eta_j^k\}_{j=1}^{S_k}$ of the random variable ξ. Compute $\hat{H}(\eta^k, y_{k+0.5}) = \frac{1}{S_k}\sum_{j=1}^{S_k} T(\eta_j^k, y_{k+0.5})$, $y_{k+1} = P_\Omega\left(x_k - t_k \hat{H}(\eta^k, y_{k+0.5})\right)$ and $\tilde{F}_{t_k}(\eta^k, x_k) = y_{k+1} - x_k$.

If $\|\tilde{F}_{t_k}(\eta^k, x_k)\| < \min\{\|F_{t_k}(\xi^k, x_k)\|, \nu\sigma_k^{-\tau}\}$, set

$$\alpha_k(\xi^k, \eta^k) = \frac{\langle \tilde{F}_{t_k}(\eta^k, x_k), \tilde{F}_{t_k}(\eta^k, x_k) - F_{t_k}(\xi^k, x_k)\rangle}{\|\tilde{F}_{t_k}(\eta^k, x_k) - F_{t_k}(\xi^k, x_k)\|^2}. \tag{4.1.7}$$

Otherwise, set $\alpha_k(\xi^k, \eta^k) = M + 1$.

Step 4: If $|\alpha_k(\xi^k, \eta^k)| \leq M$, set

$$x_{k+1} = \alpha_k(\xi^k, \eta^k)x_k + (1 - \alpha_k(\xi^k, \eta^k))y_{k+1}, \quad \sigma_{k+1} = \sigma_k + 1. \tag{4.1.8}$$

Otherwise, set

$$x_{k+1} = y_{k+1}, \quad \sigma_{k+1} = \sigma_k. \tag{4.1.9}$$

end for

Remark 4.2. *Denote* $L := \mathbb{E}[L(\xi)]$ *and* $L_p := (\mathbb{E}[L(\xi)^p])^{\frac{1}{p}} + L$ *for any* $p \geq 1$. *If Assumption 4.2 holds, then*

(i) by Jensen's inequality, for any $x, y \in \mathbb{R}^n$, *we obtain*

$$\begin{aligned}&\|\mathbb{E}[T(\xi,x)] - \mathbb{E}[T(\xi,y)]\| = \|\mathbb{E}[T(\xi,x) - T(\xi,y)]\| \\ \leq &\mathbb{E}[\|T(\xi,x) - T(\xi,y)\|] \leq \mathbb{E}[L(\xi)\|x-y\|] = L\|x-y\|,\end{aligned}$$

which means that $\mathbb{E}[T(\xi,\cdot)]$ *is L-Lipschitz continuous;*

(ii) by Minkowski's inequality and (i), for any $x, x_* \in \mathbb{R}^n$, *we have*

$$\begin{aligned}&(\mathbb{E}[\|\epsilon(\xi,x)\|^p])^{\frac{1}{p}} \\ =&(\mathbb{E}[\|T(\xi,x) - \mathbb{E}[T(\xi,x)]\|^p])^{\frac{1}{p}} \\ \leq&(\mathbb{E}[\|T(\xi,x) - T(\xi,x_*)\|^p])^{\frac{1}{p}} + (\mathbb{E}[\|T(\xi,x_*) - \mathbb{E}[T(\xi,x_*)]\|^p])^{\frac{1}{p}} \\ &+ \|\mathbb{E}[T(\xi,x)] - \mathbb{E}[T(\xi,x_*)]\| \\ \leq&(\mathbb{E}[L(\xi)^p])^{\frac{1}{p}}\|x-x_*\| + (\mathbb{E}[\|\epsilon(\xi,x_*)\|^p])^{\frac{1}{p}} + L\|x-x_*\| \\ =&L_p\|x-x_*\| + (\mathbb{E}[\|\epsilon(\xi,x_*)\|^p])^{\frac{1}{p}}, \quad \forall p \geq 1.\end{aligned} \tag{4.1.10}$$

Define

$$\mathcal{F}_0 = \sigma(x_0) \quad \text{and} \quad \hat{\mathcal{F}}_0 = \sigma(x_0, \xi^0),$$

and for $k \geq 1$,

$$\mathcal{F}_k = \sigma(x_0, \xi^0, \cdots, \xi^{k-1}, \eta^0, \cdots, \eta^{k-1}) \quad \text{and} \quad \hat{\mathcal{F}}_k = \sigma(x_0, \xi^0, \cdots, \xi^k, \eta^0, \cdots, \eta^{k-1})$$

as the σ-algebras related to the generation of $y_{k+0.5}$, y_k and x_k. It can be observed that $\mathcal{F}_k \subset \hat{\mathcal{F}}_k$, $x_k, y_k \in \mathcal{F}_k, y_{k+0.5} \in \hat{\mathcal{F}}_k$ and $y_{k+0.5} \notin \mathcal{F}_k$. Recalling (4.1.3)-(4.1.4) and the stochastic EG-Anderson(1) algorithm, we define the oracle errors by

$$\begin{cases}\hat{\epsilon}_1^k := \hat{\epsilon}(\xi^k, x_k) = \hat{H}(\xi^k, x_k) - \mathbb{E}[T(\xi, x_k)], \\ \hat{\epsilon}_2^k := \hat{\epsilon}(\eta^k, y_{k+0.5}) = \hat{H}(\eta^k, y_{k+0.5}) - \mathbb{E}[T(\xi, y_{k+0.5})], \\ \hat{\epsilon}_3^k := \hat{\epsilon}(\xi^k, y_{k+0.5}) = \hat{H}(\xi^k, y_{k+0.5}) - \mathbb{E}[T(\xi, y_{k+0.5})].\end{cases} \tag{4.1.11}$$

Then $y_{k+0.5}$ and y_{k+1} in the stochastic EG-Anderson(1) algorithm can be written as follows:

$$\begin{cases} y_{k+0.5} = P_\Omega \left(x_k - t_k(\mathbb{E}[T(\xi, x_k)] + \hat{\epsilon}_1^k) \right), \\ y_{k+1} = P_\Omega \left(x_k - t_k(\mathbb{E}[T(\xi, y_{k+0.5})] + \hat{\epsilon}_2^k) \right). \end{cases} \tag{4.1.12}$$

Remark 4.3. *Since ξ^k is independent of $\mathcal{F}_k$ and $x_k \in \mathcal{F}_k$, we have*

$$\begin{aligned} \mathbb{E}[\hat{\epsilon}_1^k|\mathcal{F}_k] =& \mathbb{E}\left[\frac{1}{S_k} \sum_{j=1}^{S_k} T(\xi_j^k, x_k) - \mathbb{E}[T(\xi, x_k)] | \mathcal{F}_k \right] \\ =& \frac{1}{S_k} \sum_{j=1}^{S_k} \mathbb{E}[T(\xi_j^k, x_k)|\mathcal{F}_k] - \mathbb{E}[T(\xi, x_k)] \\ =& \frac{1}{S_k} \sum_{j=1}^{S_k} \mathbb{E}[T(\xi, x_k)] - \mathbb{E}[T(\xi, x_k)] = \mathbf{0}. \end{aligned}$$

Similarly, from the fact that η^k is independent of $\hat{\mathcal{F}}_k$ and $y_{k+0.5} \in \hat{\mathcal{F}}_k$, we obtain $\mathbb{E}[\hat{\epsilon}_2^k|\hat{\mathcal{F}}_k] = \mathbf{0}$. In view of $\mathcal{F}_k \subset \hat{\mathcal{F}}_k$, we conclude that $\mathbb{E}[\hat{\epsilon}_2^k|\mathcal{F}_k] = \mathbb{E}\left[\mathbb{E}[\hat{\epsilon}_2^k|\hat{\mathcal{F}}_k]|\mathcal{F}_k\right] = \mathbf{0}$. Thus, $\{\hat{\epsilon}_1^k\}$ and $\{\hat{\epsilon}_2^k\}$ are the martingale difference sequences with respect to $\mathcal{F}_k$.

The following result on super-martingale convergence will be utilized, serving as a key instrument in demonstrating the convergence of SA methods.

Theorem 4.1. *[58] Let $(X, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots$ a sequence of sub-σ-algebras of $\mathcal{F}$. For each $k = 0, 1, \ldots$, let $\{y_k\}, \{u_k\}, \{a_k\}, \{b_k\}$ be four sequences of random variables. Suppose*

(i) the random variables $\{y_k\}, \{u_k\}, \{a_k\}$ and $\{b_k\}$ are nonnegative $\mathcal{F}_k$-measurable;

(ii) a.s. for each k,

$$\mathbb{E}[y_{k+1}|\mathcal{F}_k] \leq (1 + a_k)y_k - u_k + b_k;$$

(iii) a.s. $\sum_{k=0}^{\infty} a_k < \infty$, $\sum_{k=0}^{\infty} b_k < \infty$.

Then, a.s. $\{y_k\}$ converges and $\sum_{k=0}^{\infty} u_k < \infty$.

4.1.2 Convergence analysis

In this section, we will present some convergence properties of the stochastic EG-Anderson(1) algorithm. Similar to EG-Anderson(1) algorithm, we divide the iterations of the stochastic EG-Anderson(1) algorithm into three subsets

$$K_{SRE} = \{u_0, u_1, \cdots\}, \quad K_{SAA} = \{k_0, k_1, \cdots\} \quad \text{and} \quad K_{SEG} = \{l_0, l_1, \cdots\},$$

where K_{SRE} consists of the iterations defined by (4.1.5), K_{SAA} comprises iterations defined by (4.1.8), and K_{SEG} contains the remaining iterations defined by (4.1.9).

We begin by proving that the stochastic EG-Anderson(1) algorithm is well-defined.

Lemma 4.2. *Suppose that Assumption 4.2 holds. For each iteration k where line search is performed, the line search* (4.1.6) *in the stochastic EG-Anderson(1) algorithm terminates in a finite number of iterations.*

Proof In the stochastic EG-Anderson(1) algorithm, the updated form of t_k can be re-expressed as $t_k = \gamma\rho^{m_k}$, where m_k is defined as the smallest nonnegative integer m that satisfies the following inequality:

$$\gamma\rho^m\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5}^{(m)})\| \leq \mu\|x_k - y_{k+0.5}^{(m)}\|,$$

where $y_{k+0.5}^{(m)} = P_\Omega\left(x_k - \gamma\rho^m\hat{H}(\xi^k, x_k)\right)$. We prove the lemma by contradiction, assuming that the line search (4.1.6) does not terminate within a finite number of iterations. This means that, for every m, we have

$$\gamma\rho^m\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5}^{(m)})\| > \mu\|x_k - y_{k+0.5}^{(m)}\|. \tag{4.1.13}$$

We will discuss it in two cases.

(i) If $x_k \in \Omega$, then $\lim_{m\to\infty}\|y_{k+0.5}^{(m)} - x_k\| = 0$. Combining Assumption 4.2, we deduce that

$$\lim_{m\to\infty}\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5}^{(m)})\| = 0. \tag{4.1.14}$$

From (4.1.13) and Lemma 2.2, we have

$$\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5}^{(m)})\| > \frac{\mu}{\gamma\rho^m}\|x_k - y_{k+0.5}^{(m)}\| \geq \frac{\mu}{\gamma}\|F_\gamma(\xi^k, x_k)\| > 0.$$

Taking the limit $m \to \infty$ in the above inequality, we can find a contradiction with (4.1.14).

(ii) If $x_k \notin \Omega$, then

$$\lim_{m\to\infty} \|x_k - y_{k+0.5}^{(m)}\| > 0, \tag{4.1.15}$$

and

$$\lim_{m\to\infty} \gamma\rho^m \|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5}^{(m)})\| = 0. \tag{4.1.16}$$

Letting $m \to \infty$ in (4.1.13), we obtain a contradiction with (4.1.15) and (4.1.16).

Thus, the line search (4.1.6) is well-defined. □

The stochastic EG-Anderson(1) algorithm employs a dynamic sampled SA line search method to address the lack of Lipschitz constants. The following lemma provides a lower bound for the step size.

Lemma 4.3. *Suppose that Assumptions 4.2 and 4.3 hold. Let $\{t_k\}$ be the step-size sequence generated by the stochastic EG-Anderson(1) algorithm and $\hat{L}_k(\xi^k) := \frac{1}{S_k}\sum_{j=1}^{S_k} L(\xi_j^k)$. Then we have that a.s. $t_k\hat{L}_k(\xi^k) \geq \min\{\rho\mu, \gamma\}$ and $(\mathbb{E}[t_k^2|\mathcal{F}_k])^{\frac{1}{2}} \cdot (\mathbb{E}[L(\xi)^2])^{\frac{1}{2}} \geq \min\{\rho\mu, \gamma\}$.*

Proof We first discuss the lower bound of t_k by dividing it into two cases.

(i) If line search (4.1.6) is not performed, or γ satisfies (4.1.6) during the line search, then $t_k = \gamma$.

(ii) Otherwise, $t_k < \gamma$. Let $\hat{y}_{k+0.5} := P_\Omega(x_k - t_k\rho^{-1}\hat{H}(\xi^k, x_k))$. By $t_k < \gamma$, $t_k < t_k\rho^{-1}$ and Lemma 2.2, we have

$$\|y_{k+0.5} - x_k\| \geq \frac{t_k}{\gamma}\|F_\gamma(\xi^k, x_k)\| > 0$$

and

$$\|\hat{y}_{k+0.5} - x_k\| \geq \|y_{k+0.5} - x_k\| > 0.$$

Then we find

$$t_k\rho^{-1}\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, \hat{y}_{k+0.5})\| > \mu\|x_k - \hat{y}_{k+0.5}\| > 0. \tag{4.1.17}$$

By Assumption 4.2-(i), we have

$$\begin{aligned}\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, \hat{y}_{k+0.5})\| = &\left\|\frac{1}{S_k}\sum_{j=1}^{S_k} T(\xi_j^k, x_k) - \frac{1}{S_k}\sum_{j=1}^{S_k} T(\xi_j^k, \hat{y}_{k+0.5})\right\| \\ \leq &\hat{L}_k(\xi^k)\|x_k - \hat{y}_{k+0.5}\|.\end{aligned} \tag{4.1.18}$$

Combining (4.1.17) and (4.1.18), we obtain $t_k \geq \frac{\rho\mu}{\hat{L}_k(\xi^k)}$.

Thus, we conclude that $t_k \geq \min\left\{\frac{\rho\mu}{\hat{L}_k(\xi^k)}, \gamma\right\}$.

Since a.s. $L(\xi) \geq 1$, we have that a.s. $t_k\hat{L}_k(\xi^k) \geq \min\{\rho\mu, \gamma\}$. Combining the Hölder's inequality, we deduce that

$$\begin{aligned}\min\{\rho\mu, \gamma\} \leq \mathbb{E}\left[t_k\hat{L}_k(\xi^k)|\mathcal{F}_k\right] &\leq \left(\mathbb{E}[t_k^2|\mathcal{F}_k]\right)^{\frac{1}{2}}\left(\mathbb{E}[\hat{L}_k(\xi^k)^2|\mathcal{F}_k]\right)^{\frac{1}{2}} \\ &\leq \left(\mathbb{E}[t_k^2|\mathcal{F}_k]\right)^{\frac{1}{2}}\left(\frac{1}{S_k}\sum_{j=1}^{S_k}\mathbb{E}[L(\xi_j^k)^2|\mathcal{F}_k]\right)^{\frac{1}{2}} = \left(\mathbb{E}[t_k^2|\mathcal{F}_k]\right)^{\frac{1}{2}} \cdot \left(\mathbb{E}[L(\xi)^2]\right)^{\frac{1}{2}},\end{aligned}$$

where the last equality follows from Assumption 4.3. □

The recursive relations given by the following two lemmas play a crucial role in the main convergence results.

Lemma 4.4. *Suppose that Assumptions 4.1 and 4.2 hold. Let* $\{x_k\}$, $\{y_{k+0.5}\}$ *and* $\{y_{k+1}\}$ *be the sequences generated by the stochastic EG-Anderson(1) algorithm. Then for all* $x^* \in \mathrm{SSOL}\,(\Omega, \mathbb{E}[T(\xi,\cdot)])$, *we have*

$$\begin{aligned}\|y_{k+1}-x^*\|^2 \leq & \|x_k-x^*\|^2 - \frac{(1-3\mu^2)t_k^2}{2\gamma^2} r_\gamma(x_k)^2 + (1-3\mu^2)\gamma^2\|\hat{\epsilon}_1^k\|^2 \\ & + 3\gamma^2\|\hat{\epsilon}_2^k\|^2 + 3\gamma^2\|\hat{\epsilon}_3^k\|^2 + 2\langle t_k\hat{\epsilon}_2^k, x^*-y_{k+0.5}\rangle.\end{aligned} \tag{4.1.19}$$

Proof By Assumption 4.1 and $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi,\cdot)])$, we have

$$\langle \mathbb{E}[T(\xi, y_{k+0.5})], y_{k+0.5}-x^*\rangle \geq 0,$$

which implies

$$\langle t_k\mathbb{E}[T(\xi, y_{k+0.5})], x^*-y_{k+1}\rangle \leq \langle t_k\mathbb{E}[T(\xi, y_{k+0.5})], y_{k+0.5}-y_{k+1}\rangle\,. \tag{4.1.20}$$

By the definition of $y_{k+0.5}$ and Lemma 2.1-(iii), we find

$$\langle x_k - t_k\hat{H}(\xi^k, x_k) - y_{k+0.5}, y_{k+1}-y_{k+0.5}\rangle \leq 0.$$

This implies

$$\begin{aligned}&\langle x_k - t_k\mathbb{E}[T(\xi, y_{k+0.5})] - y_{k+0.5}, y_{k+1}-y_{k+0.5}\rangle \\ \leq &\langle t_k\hat{H}(\xi^k, x_k) - t_k\mathbb{E}[T(\xi, y_{k+0.5})], y_{k+1}-y_{k+0.5}\rangle.\end{aligned} \tag{4.1.21}$$

Using (4.1.20) and (4.1.21), and following a similar approach to the estimation in (3.1.18), we have

$$\begin{aligned}\|y_{k+1}-x^*\|^2 \leq & \|x_k-x^*\|^2 - \|x_k-y_{k+1}\|^2 + 2\langle t_k\hat{H}(\eta^k, y_{k+0.5}), x^*-y_{k+1}\rangle \\ = & \|x_k-x^*\|^2 - \|x_k-y_{k+1}\|^2 + 2\langle t_k\mathbb{E}[T(\xi, y_{k+0.5})], x^*-y_{k+1}\rangle \\ & + 2\langle t_k\hat{\epsilon}_2^k, x^*-y_{k+1}\rangle \\ \overset{(4.1.20)}{\leq} & \|x_k-x^*\|^2 - \|x_k-y_{k+1}\|^2 + 2\langle t_k\mathbb{E}[T(\xi, y_{k+0.5})], y_{k+0.5}-y_{k+1}\rangle \\ & + 2\langle t_k\hat{\epsilon}_2^k, x^*-y_{k+1}\rangle\end{aligned}$$

$$
\begin{aligned}
&= \|x_k - x^*\|^2 - \|x_k - y_{k+0.5}\|^2 - \|y_{k+0.5} - y_{k+1}\|^2 \\
&\quad + 2\langle x_k - t_k\mathbb{E}[T(\xi, y_{k+0.5})] - y_{k+0.5}, y_{k+1} - y_{k+0.5}\rangle \\
&\quad + 2\langle t_k\hat{\epsilon}_2^k, x^* - y_{k+1}\rangle \\
&\overset{(4.1.21)}{\leq} \|x_k - x^*\|^2 - \|x_k - y_{k+0.5}\|^2 - \|y_{k+0.5} - y_{k+1}\|^2 \\
&\quad + 2\langle t_k\hat{H}(\xi^k, x_k) - t_k\mathbb{E}[T(\xi, y_{k+0.5})], y_{k+1} - y_{k+0.5}\rangle \\
&\quad + 2\langle t_k\hat{\epsilon}_2^k, x^* - y_{k+1}\rangle \\
&= \|x_k - x^*\|^2 - \|x_k - y_{k+0.5}\|^2 - \|y_{k+0.5} - y_{k+1}\|^2 \\
&\quad + 2\langle t_k\hat{H}(\xi^k, x_k) - t_k\hat{H}(\eta^k, y_{k+0.5}), y_{k+1} - y_{k+0.5}\rangle \\
&\quad + 2\langle t_k\hat{\epsilon}_2^k, x^* - y_{k+0.5}\rangle. \qquad (4.1.22)
\end{aligned}
$$

Next, we estimate the term $2\langle t_k\hat{H}(\xi^k, x_k) - t_k\hat{H}(\eta^k, y_{k+0.5}), y_{k+1} - y_{k+0.5}\rangle$ in (4.1.22). By the Cauchy-Schwarz inequality and the Young's inequality, we have

$$
\begin{aligned}
&2\langle t_k\hat{H}(\xi^k, x_k) - t_k\hat{H}(\eta^k, y_{k+0.5}), y_{k+1} - y_{k+0.5}\rangle \\
\leq\ & t_k^2\|\hat{H}(\xi^k, x_k) - \hat{H}(\eta^k, y_{k+0.5})\|^2 + \|y_{k+1} - y_{k+0.5}\|^2 \\
\leq\ & 3t_k^2\|\hat{H}(\xi^k, x_k) - \hat{H}(\xi^k, y_{k+0.5})\|^2 + 3\gamma^2\|\hat{\epsilon}_2^k\|^2 + 3\gamma^2\|\hat{\epsilon}_3^k\|^2 \qquad (4.1.23) \\
&+ \|y_{k+1} - y_{k+0.5}\|^2 \\
\overset{(4.1.6)}{\leq}\ & 3\mu^2\|x_k - y_{k+0.5}\|^2 + 3\gamma^2\|\hat{\epsilon}_2^k\|^2 + 3\gamma^2\|\hat{\epsilon}_3^k\|^2 + \|y_{k+1} - y_{k+0.5}\|^2.
\end{aligned}
$$

Substituting (4.1.23) into (4.1.22), we deduce that

$$
\begin{aligned}
\|y_{k+1} - x^*\|^2 \leq & \|x_k - x^*\|^2 - (1 - 3\mu^2)\|x_k - y_{k+0.5}\|^2 + 3\gamma^2\|\hat{\epsilon}_2^k\|^2 + 3\gamma^2\|\hat{\epsilon}_3^k\|^2 \\
&+ 2\langle t_k\hat{\epsilon}_2^k, x^* - y_{k+0.5}\rangle. \qquad (4.1.24)
\end{aligned}
$$

Recalling the function $r_t(x) := \|P_\Omega(x - t\mathbb{E}[T(\xi, x)]) - x\|$ defined in (4.1.2), and

combining this with $t_k \leq \gamma$ and Lemma 2.2, we obtain

$$\begin{aligned} t_k^2 r_\gamma(x_k)^2 \leq & \gamma^2 r_{t_k}(x_k)^2 = \gamma^2 \|P_\Omega(x_k - t_k \mathbb{E}[T(\xi, x_k)]) - x_k\|^2 \\ \leq & 2\gamma^2 \|x_k - y_{k+0.5}\|^2 + 2\gamma^2 \|P_\Omega(x_k - t_k \mathbb{E}[T(\xi, x_k)]) - P_\Omega(x_k - t_k \hat{H}(\xi^k, x_k))\|^2 \\ \leq & 2\gamma^2 \|x_k - y_{k+0.5}\|^2 + 2\gamma^4 \|\hat{\epsilon}_1^k\|^2, \end{aligned}$$

which implies

$$\|x_k - y_{k+0.5}\|^2 \geq \frac{t_k^2}{2\gamma^2} r_\gamma(x_k)^2 - \gamma^2 \|\hat{\epsilon}_1^k\|^2. \tag{4.1.25}$$

Plugging (4.1.25) into (4.1.24) yields

$$\begin{aligned} \|y_{k+1} - x^*\|^2 \leq & \|x_k - x^*\|^2 - \frac{(1-3\mu^2)t_k^2}{2\gamma^2} r_\gamma(x_k)^2 + (1-3\mu^2)\gamma^2 \|\hat{\epsilon}_1^k\|^2 \\ & + 3\gamma^2 \|\hat{\epsilon}_2^k\|^2 + 3\gamma^2 \|\hat{\epsilon}_3^k\|^2 + 2\langle t_k \hat{\epsilon}_2^k, x^* - y_{k+0.5} \rangle. \end{aligned}$$

□

Lemma 4.5. *Suppose that Assumptions 4.1 and 4.2 hold. Let $\{x_k\}$ and $\{y_{k+0.5}\}$ be the sequences generated by the stochastic EG-Anderson(1) algorithm. Then, there exists $\hat{\beta} \in (0, \frac{1}{2}]$ such that for any $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$,*

$$\begin{aligned} \|x_{k+1} - x^*\|^2 \leq & \|x_k - x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta} t_k^2}{2\gamma^2} r_\gamma(x_k)^2 + (1-3\mu^2)(M+1)\gamma^2 \|\hat{\epsilon}_1^k\|^2 \\ & + 3\gamma^2 (M+1) \|\hat{\epsilon}_2^k\|^2 + 3\gamma^2 (M+1) \|\hat{\epsilon}_3^k\|^2 \\ & + 2(M+1) \left| \langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle \right| + \varrho_k, \end{aligned} \tag{4.1.26}$$

where

$$\varrho_k = \begin{cases} M(M+1)\nu^2 (1+i)^{-2\tau}, & \text{if} \quad k \in K_{SAA} := \{k_i, i = 1, 2, \ldots\}, \\ 0, & \text{if} \quad k \in K_{SRE} \cup K_{SEG}. \end{cases} \tag{4.1.27}$$

Proof We will discuss this by dividing it into three cases.

(i) For $u_m \in K_{SRE}$, we know that $\|F_\gamma(\xi^{u_m}, x_{u_m})\| = \|P_\Omega(x_{u_m} - \gamma\hat{H}(\xi^{u_m}, x_{u_m})) - x_{u_m}\| = 0$ and $x_{u_m+1} = x_{u_m}$. Then we have

$$\begin{aligned} r_\gamma(x_{u_m})^2 =& \|P_\Omega\left(x_{u_m} - \gamma\mathbb{E}[T(\xi, x_{u_m})]\right) - x_{u_m}\|^2 \\ \leq& 2\|F_\gamma(\xi^{u_m}, x_{u_m})\|^2 + 2\|P_\Omega(x_{u_m} - \gamma\mathbb{E}[T(\xi, x_{u_m})]) \\ & - P_\Omega(x_{u_m} - \gamma\hat{H}(\xi^{u_m}, x_{u_m}))\|^2 \\ \leq& 2\gamma^2\|\hat{\epsilon}_1^{u_m}\|^2. \end{aligned}$$

Combining with $t_{u_m} = \gamma$ and $\mu \in (0, \frac{\sqrt{3}}{3})$, we obtain $-\frac{(1-3\mu^2)t_{u_m}^2}{2\gamma^2} r_\gamma(x_{u_m})^2 + (1 - 3\mu^2)\gamma^2\|\hat{\epsilon}_1^{u_m}\|^2 \geq 0$. Therefore, we conclude that

$$\begin{aligned} & \|x_{u_m+1} - x^*\|^2 \\ =& \|x_{u_m} - x^*\|^2 \\ \leq& \|x_{u_m} - x^*\|^2 - \frac{(1-3\mu^2)t_{u_m}^2}{2\gamma^2} r_\gamma(x_{u_m})^2 + (1-3\mu^2)\gamma^2\|\hat{\epsilon}_1^{u_m}\|^2 \\ & + 3\gamma^2\|\hat{\epsilon}_2^{u_m}\|^2 + 3\gamma^2\|\hat{\epsilon}_3^{u_m}\|^2 + 2\left|\langle t_{u_m}\hat{\epsilon}_2^{u_m}, x^* - y_{u_m+0.5}\rangle\right|. \end{aligned} \tag{4.1.28}$$

(ii) For $k_i \in K_{SAA}$, let $\beta_{k_i}(\xi^{k_i}, \eta^{k_i}) := 1 - \alpha_{k_i}(\xi^{k_i}, \eta^{k_i})$. By the definition of x_{k_i+1} in (4.1.8), applying a similar estimation method as in (3.1.19) yields

$$\begin{aligned} \|x_{k_i+1} - x^*\|^2 =& \alpha_{k_i}(\xi^{k_i}, \eta^{k_i})\|x_{k_i} - x^*\|^2 + \beta_{k_i}(\xi^{k_i}, \eta^{k_i})\|y_{k_i+1} - x^*\|^2 \\ & - \alpha_{k_i}(\xi^{k_i}, \eta^{k_i})\beta_{k_i}(\xi^{k_i}, \eta^{k_i})\|x_{k_i} - y_{k_i+1}\|^2. \end{aligned} \tag{4.1.29}$$

Since $\|\tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i})\| < \|F_{t_{k_i}}(\xi^{k_i}, x_{k_i})\|$ when $k_i \in K_{SAA}$, then there exists $\hat{\beta} \in (0, \frac{1}{2}]$ such that

$$\hat{\beta}\|\tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i})\|^2 + (1-2\hat{\beta})\langle F_{t_{k_i}}(\xi^{k_i}, x_{k_i}), \tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i})\rangle < (1-\hat{\beta})\|F_{t_{k_i}}(\xi^{k_i}, x_{k_i})\|^2,$$

which implies

$$\langle \tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i}) - F_{t_{k_i}}(\xi^{k_i}, x_{k_i}), \tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i})\rangle < (1-\hat{\beta})\|\tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i}) - F_{t_{k_i}}(\xi^{k_i}, x_{k_i})\|^2.$$

By (4.1.7), we have $\alpha_{k_i}(\xi^{k_i},\eta^{k_i}) < 1-\hat{\beta}$. Thus $\beta_{k_i}(\xi^{k_i},\eta^{k_i}) = 1-\alpha_{k_i}(\xi^{k_i},\eta^{k_i}) > \hat{\beta}$. Substituting (4.1.19) into (4.1.29) and using $\mu \in (0, \frac{\sqrt{3}}{3})$, $|\alpha_{k_i}(\xi^{k_i},\eta^{k_i})| \leq M$, $\hat{\beta} < \beta_{k_i}(\xi^{k_i},\eta^{k_i}) \leq M+1$, $\|x_{k_i} - y_{k_i+1}\| = \|\tilde{F}_{t_{k_i}}(\eta^{k_i}, x_{k_i})\| \leq \nu\sigma_{k_i}^{-\tau} = \nu(\sigma_0 + i)^{-\tau} = \nu(1+i)^{-\tau}$ and (4.1.27), we deduce that

$$\begin{aligned}\|x_{k_i+1} - x^*\|^2 \leq & \|x_{k_i} - x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta}t_{k_i}^2}{2\gamma^2} r_\gamma(x_{k_i})^2 + (1-3\mu^2)(M+1)\gamma^2\|\hat{\epsilon}_1^{k_i}\|^2 \\ & + 3\gamma^2(M+1)\|\hat{\epsilon}_2^{k_i}\|^2 + 3\gamma^2(M+1)\|\hat{\epsilon}_3^{k_i}\|^2 \\ & + 2(M+1)\left|\langle x^* - y_{k_i+0.5}, t_{k_i}\hat{\epsilon}_2^{k_i}\rangle\right| + \varrho_{k_i}.\end{aligned} \tag{4.1.30}$$

(iii) For $l_j \in K_{SEG}$, by Lemma 4.4, we deduce that

$$\begin{aligned}\|x_{l_j+1} - x^*\|^2 = & \|y_{l_j+1} - x^*\|^2 \\ \leq & \|x_{l_j} - x^*\|^2 - \frac{(1-3\mu^2)t_{l_j}^2}{2\gamma^2} r_\gamma(x_{l_j})^2 + (1-3\mu^2)\gamma^2\|\hat{\epsilon}_1^{l_j}\|^2 \\ & + 3\gamma^2\|\hat{\epsilon}_2^{l_j}\|^2 + 3\gamma^2\|\hat{\epsilon}_3^{l_j}\|^2 + 2\left|\langle t_{l_j}\hat{\epsilon}_2^{l_j}, x^* - y_{l_j+0.5}\rangle\right|.\end{aligned} \tag{4.1.31}$$

Since $\varrho_{l_j} = \varrho_{u_m} = 0$ in (4.1.27), and by combining (4.1.28), (4.1.30) and (4.1.31), for every k, we find

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \leq & \|x_k - x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta}t_k^2}{2\gamma^2} r_\gamma(x_k)^2 + (1-3\mu^2)(M+1)\gamma^2\|\hat{\epsilon}_1^k\|^2 \\ & + 3\gamma^2(M+1)\|\hat{\epsilon}_2^k\|^2 + 3\gamma^2(M+1)\|\hat{\epsilon}_3^k\|^2 \\ & + 2(M+1)\left|\langle x^* - y_{k+0.5}, t_k\hat{\epsilon}_2^k\rangle\right| + \varrho_k,\end{aligned}$$

where $\sum_{k=0}^{\infty}\varrho_k = \sum_{i=0}^{\infty}\varrho_{k_i} < \infty$ holds because $\tau > \frac{1}{2}$. □

Lemma 4.5 indicates that both bounding $\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k\hat{\epsilon}_2^k\rangle\right| | \mathcal{F}_k\right]$ and the second order moments of the sequences $\{\hat{\epsilon}_1^k\}$, $\{\hat{\epsilon}_2^k\}$ and $\{\hat{\epsilon}_3^k\}$ are essential for establishing the convergence properties of the stochastic EG-Anderson(1) algorithm.

The following lemma is mainly used to control the oracle errors given above.

Lemma 4.6 (Burkholder-Davis-Gundy inequality)**.** *[11] For any $p \geq 2$ and $N \in \mathbb{N}$, there exists $B_p > 0$ such that for any vector-valued martingale $(y_i, \mathcal{F}_i)_{i=1}^N$ taking values in $\mathbb{R}^n$ with $y_0 = 0$, it holds that*

$$(\mathbb{E}[\|y_N\|^p])^{\frac{1}{p}} \leq \left(\mathbb{E}\left[\left(\sup_{i\leq N}\|y_i\|\right)^p\right]\right)^{\frac{1}{p}} \leq B_p \left(\mathbb{E}\left[\left(\sum_{i=1}^N \|y_i - y_{i-1}\|^2\right)^{\frac{p}{2}}\right]\right)^{\frac{1}{p}}$$

$$\leq B_p \sqrt{\sum_{i=1}^N (\mathbb{E}[\|y_i - y_{i-1}\|^p])^{\frac{2}{p}}}.$$

Recalling the definitions of $\epsilon(\xi, x)$ in (4.1.3) and $\hat{\epsilon}(\xi^N, x)$ in (4.1.4), respectively, we know that

$$\hat{\epsilon}(\xi^N, x) := \frac{1}{N}\sum_{j=1}^N \left(T(\xi_j, x) - \mathbb{E}[T(\xi, x)]\right).$$

In [37], the estimation of $\left(\mathbb{E}[\|\hat{\epsilon}(\xi^N, x)\|^p]\right)^{\frac{1}{p}}$ is provided under certain conditions. To present the conclusions of this thesis more clearly, Lemma 4.7 gives a proof similar to the method used in [37].

Lemma 4.7. *Suppose that Assumption 4.2 holds. Given $N \in \mathbb{N}$, let $\xi^N := \{\xi_j\}_{j=1}^N$ be an i.i.d. sample of ξ. Then, for any $p \geq 2$, $x, x_* \in \mathbb{R}^n$, we have*

$$\left(\mathbb{E}[\|\hat{\epsilon}(\xi^N, x)\|^p]\right)^{\frac{1}{p}} \leq B_p \frac{\delta_p(x_*)\left(1 + \|x - x_*\|\right)}{\sqrt{N}}$$

and

$$\left(\mathbb{E}\left[\left|\langle v, \hat{\epsilon}(\xi^N, x)\rangle\right|^p\right]\right)^{\frac{1}{p}} \leq \|v\| B_p \frac{\delta_p(x_*)\left(1 + \|x - x_*\|\right)}{\sqrt{N}}, \quad \forall v \in \mathbb{R}^n,$$

where B_p is the same as in Lemma 4.6 and $\delta_p(x_*) = \max\left\{(\mathbb{E}[\|\epsilon(\xi, x_*)\|^p])^{\frac{1}{p}}, L_p\right\}$ with L_p defined in Remark 4.2.

Proof Given $N \in \mathbb{N}$, we define the sequence $\{U^t\}_{t=0}^N$ by setting $U^0 = 0$ and $U^t := \sum_{j=1}^t \frac{T(\xi_j, x) - \mathbb{E}[T(\xi, x)]}{N}$ for $1 \leq t \leq N$. We also define $\mathcal{G}_0 := \sigma(U^0)$ and

$\mathcal{G}_t := \sigma(\xi_1, \ldots, \xi_t)$ for $1 \leq t \leq N$. Since $\{\xi_j\}_{j=1}^N$ is an i.i.d. sample drawn from Ξ, $\{U^t, \mathcal{G}_t\}_{t=0}^N$ is an $\mathbb{R}^n$-valued martingale. By Assumption 4.2 and (4.1.10), for any $p \geq 2$ and $x_* \in \mathbb{R}^n$, we deduce

$$\begin{aligned}\left(\mathbb{E}\left[\|U^t - U^{t-1}\|^p\right]\right)^{\frac{1}{p}} &= \left(\mathbb{E}\left[\left(\frac{\|T(\xi, x) - \mathbb{E}[T(\xi, x)]\|}{N}\right)^p\right]\right)^{\frac{1}{p}} \\ &\leq \frac{\left(\mathbb{E}\left[\|\epsilon(\xi, x_*)\|^p\right]\right)^{\frac{1}{p}} + L_p\|x - x_*\|}{N}.\end{aligned}$$

Then from Lemma 4.6, there exists $B_p > 0$ such that

$$\begin{aligned}\left(\mathbb{E}\left[\|\hat{\epsilon}(\xi^N, x)\|^p\right]\right)^{\frac{1}{p}} = \left(\mathbb{E}\left[\|U^N\|^p\right]\right)^{\frac{1}{p}} &\leq B_p\sqrt{\sum_{t=1}^N (\mathbb{E}[\|U^t - U^{t-1}\|^p])^{\frac{2}{p}}} \\ &\leq B_p\frac{\left(\mathbb{E}[\|\epsilon(\xi, x_*)\|^p]\right)^{\frac{1}{p}} + L_p\|x - x_*\|}{\sqrt{N}} \\ &\leq B_p\frac{\delta_p(x_*)\left(1 + \|x - x_*\|\right)}{\sqrt{N}}.\end{aligned}$$

Moreover, for any $v \in \mathbb{R}^n$, we find that

$$\begin{aligned}\left(\mathbb{E}\left[\left|\langle v, \hat{\epsilon}(\xi^N, x)\rangle\right|^p\right]\right)^{\frac{1}{p}} &\leq \|v\|\left(\mathbb{E}\left[\|\hat{\epsilon}(\xi^N, x)\|^p\right]\right)^{\frac{1}{p}} \\ &\leq \|v\|B_p\frac{\delta_p(x_*)\left(1 + \|x - x_*\|\right)}{\sqrt{N}}.\end{aligned} \tag{4.1.32}$$

□

The following lemma establishes an upper bound on the second order moment of the martingale difference sequences $\{\hat{\epsilon}_1^k\}$ and $\{\hat{\epsilon}_2^k\}$. While prior works [37, 38] prove similar bounds using the Burkholder-Davis-Gundy inequality, we present a more direct characterization.

For any $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$, we define

$$D_{k,p}(x^*) := \gamma\frac{1}{\sqrt{S_k}}B_p\delta_p(x^*) \quad \text{and} \quad I_{k,p}(x^*) := 1 + \gamma L + D_{k,p}(x^*). \tag{4.1.33}$$

Lemma 4.8. *Suppose that Assumptions 4.2 and 4.3 hold. Then for any k, $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi,\cdot)])$, $p = 2q$ and $q \geq 1$, it holds that*

(i) $\left(\mathbb{E}\left[\|\hat{\epsilon}_1^k\|^{2q}|\mathcal{F}_k\right]\right)^{\frac{1}{q}} \leq 2B_p^2 \dfrac{\delta_p(x^*)^2\left(1+\|x_k-x^*\|^2\right)}{S_k}$;

(ii) $\left(\mathbb{E}\left[\|\hat{\epsilon}_2^k\|^{2q}|\mathcal{F}_k\right]\right)^{\frac{1}{q}} \leq \dfrac{2}{S_k} B_p^2 \delta_p(x^*)^2\left(1+2I_{k,p}(x^*)^2\|x_k-x^*\|^2+2D_{k,p}(x^*)^2\right)$;

(iii) $\left(\mathbb{E}\left[\left|\langle x^*-y_{k+0.5}, t_k\hat{\epsilon}_2^k\rangle\right|^q|\mathcal{F}_k\right]\right)^{\frac{1}{q}} \leq \left(\frac{1}{2}D_{k,p}(x^*)+2D_{k,p}(x^*)I_{k,p}(x^*)^2\right)\|x_k-x^*\|^2+ \frac{1}{2}D_{k,p}(x^*)I_{k,p}(x^*)^2+D_{k,p}(x^*)^2+2D_{k,p}(x^*)^3$.

Proof We will prove the above three statements separately.

(i) Using Lemma 4.7, $x_k \in \mathcal{F}_k$, and the independence of ξ^k with $\mathcal{F}_k$, we get

$$\left(\mathbb{E}\left[\|\hat{\epsilon}_1^k\|^p|\mathcal{F}_k\right]\right)^{\frac{1}{p}} = \left(\mathbb{E}\left[\|\hat{\epsilon}_1^k\|^p\right]\right)^{\frac{1}{p}} \leq B_p \frac{\delta_p(x^*)\left(1+\|x_k-x^*\|\right)}{\sqrt{S_k}}, \tag{4.1.34}$$

which implies

$$\begin{aligned}\left(\mathbb{E}\left[\|\hat{\epsilon}_1^k\|^{2q}|\mathcal{F}_k\right]\right)^{\frac{1}{q}} &= \left(\mathbb{E}\left[\|\hat{\epsilon}_1^k\|^p|\mathcal{F}_k\right]\right)^{\frac{2}{p}} \\ &= \left(\mathbb{E}\left[\|\hat{\epsilon}_1^k\|^p\right]\right)^{\frac{2}{p}} \leq 2B_p^2 \frac{\delta_p(x^*)^2\left(1+\|x_k-x^*\|^2\right)}{S_k}.\end{aligned}$$

(ii) Recall that $y_{k+0.5} := P_\Omega\left(x_k - t_k\left(\mathbb{E}[T(\xi,x_k)]+\hat{\epsilon}_1^k\right)\right)$ in (4.1.12). By Lemma 4.1 and $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi,\cdot)])$, we have $x^* = P_\Omega\left(x^* - t_k\mathbb{E}[T(\xi,x^*)]\right)$. Then, combining Lemma 2.1-(i) with Remark 4.2-(i) and $t_k \leq \gamma$, we conclude that

$$\begin{aligned}\|x^*-y_{k+0.5}\| &= \left\|P_\Omega(x^*-t_k\mathbb{E}[T(\xi,x^*)]) - P_\Omega\left(x_k - t_k(\mathbb{E}[T(\xi,x_k)]+\hat{\epsilon}_1^k)\right)\right\| \\ &\leq \|x_k-x^*\| + t_k\|\mathbb{E}[T(\xi,x_k)]-\mathbb{E}[T(\xi,x^*)]\| + t_k\|\hat{\epsilon}_1^k\| \\ &\leq \|x_k-x^*\| + \gamma L\|x_k-x^*\| + \gamma\|\hat{\epsilon}_1^k\|,\end{aligned} \tag{4.1.35}$$

where $L = \mathbb{E}[L(\xi)]$. Taking $(\mathbb{E}[|\cdot|^p|\mathcal{F}_k])^{\frac{1}{p}}$ in (4.1.35) and using Minkowski's inequality, $x_k \in \mathcal{F}_k$, and (4.1.34), we deduce that

$$\begin{aligned}
&\left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^p|\mathcal{F}_k\right]\right)^{\frac{1}{p}} \\
\leq &\left(\mathbb{E}\left[\left(\|x_k - x^*\| + \gamma L\|x_k - x^*\| + \gamma\|\hat{\epsilon}_1^k\|\right)^p|\mathcal{F}_k\right]\right)^{\frac{1}{p}} \\
\leq &(1+\gamma L)\|x_k - x^*\| + \gamma\left(\mathbb{E}[\|\hat{\epsilon}_1^k\|^p|\mathcal{F}_k]\right)^{\frac{1}{p}} \\
\leq &(1+\gamma L)\|x_k - x^*\| + \gamma B_p\frac{\delta_p(x^*)\left(1+\|x_k - x^*\|\right)}{\sqrt{S_k}} \\
= &\left(1+\gamma L + \gamma\frac{1}{\sqrt{S_k}}B_p\delta_p(x^*)\right)\|x_k - x^*\| + \gamma\frac{1}{\sqrt{S_k}}B_p\delta_p(x^*) \\
= &I_{k,p}(x^*)\|x_k - x^*\| + D_{k,p}(x^*).
\end{aligned} \tag{4.1.36}$$

Then, we have

$$\left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^p|\mathcal{F}_k\right]\right)^{\frac{2}{p}} \leq 2I_{k,p}(x^*)^2\|x_k - x^*\|^2 + 2D_{k,p}(x^*)^2. \tag{4.1.37}$$

Since η^k is independent of $\hat{\mathcal{F}}_k$ and $y_{k+0.5} \in \hat{\mathcal{F}}_k$, Lemma 4.7 implies that

$$\left(\mathbb{E}\left[\|\hat{\epsilon}_2^k\|^p|\hat{\mathcal{F}}_k\right]\right)^{\frac{1}{p}} = \left(\mathbb{E}\left[\|\hat{\epsilon}_2^k\|^p\right]\right)^{\frac{1}{p}} \leq B_p\frac{\delta_p(x^*)\left(1+\|y_{k+0.5} - x^*\|\right)}{\sqrt{S_k}}. \tag{4.1.38}$$

Combining with Minkowski's inequality, (4.1.37), (4.1.38), $\left(\mathbb{E}[|\cdot|^q|\hat{\mathcal{F}}_k]\right)^{\frac{1}{q}} \leq \left(\mathbb{E}[|\cdot|^p|\hat{\mathcal{F}}_k]\right)^{\frac{1}{p}}$ and $\left(\mathbb{E}\left[\left(\mathbb{E}[|\cdot|^q|\hat{\mathcal{F}}_k]\right)|\mathcal{F}_k\right]\right)^{\frac{1}{q}} = \left(\mathbb{E}[|\cdot|^q|\mathcal{F}_k]\right)^{\frac{1}{q}}$, we deduce that

$$\begin{aligned}
\left(\mathbb{E}[\|\hat{\epsilon}_2^k\|^{2q}|\mathcal{F}_k]\right)^{\frac{1}{q}} = &\left(\mathbb{E}[\|\hat{\epsilon}_2^k\|^p|\mathcal{F}_k]\right)^{\frac{2}{p}} = \left(\mathbb{E}\left[\left(\mathbb{E}[\|\hat{\epsilon}_2^k\|^p|\hat{\mathcal{F}}_k]\right)|\mathcal{F}_k\right]\right)^{\frac{2}{p}} \\
&\overset{(4.1.38)}{\leq} \left(\mathbb{E}\left[\left(\frac{1}{\sqrt{S_k}}B_p\delta_p(x^*)\left(1+\|y_{k+0.5} - x^*\|\right)\right)^p|\mathcal{F}_k\right]^{\frac{1}{p}}\right)^2 \\
&\leq \left(\frac{1}{\sqrt{S_k}}B_p\delta_p(x^*) + \frac{1}{\sqrt{S_k}}B_p\delta_p(x^*)\mathbb{E}\left[\|y_{k+0.5} - x^*\|^p|\mathcal{F}_k\right]^{\frac{1}{p}}\right)^2
\end{aligned}$$

$$\leq \frac{2}{S_k} B_p^2 \delta_p(x^*)^2 \left(1 + \left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^p | \mathcal{F}_k\right]\right)^{\frac{2}{p}}\right)$$

$$\overset{(4.1.37)}{\leq} \frac{2}{S_k} B_p^2 \delta_p(x^*)^2 \left(1 + 2I_{k,p}(x^*)^2 \|x_k - x^*\|^2 + 2D_{k,p}(x^*)^2\right).$$

(iii) By Lemma 4.7, $y_{k+0.5} \in \hat{\mathcal{F}}_k$, (4.1.38) and $t_k \leq \gamma$, we deduce that

$$\begin{aligned}
&\left(\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle\right|^p | \hat{\mathcal{F}}_k\right]\right)^{\frac{1}{p}} \\
\leq\ & \gamma \|x^* - y_{k+0.5}\| \left(\mathbb{E}\left[\|\hat{\epsilon}_2^k\|^p | \hat{\mathcal{F}}_k\right]\right)^{\frac{1}{p}} \\
\overset{(4.1.38)}{\leq}\ & \gamma \|x^* - y_{k+0.5}\| B_p \frac{\delta_p(x^*)\left(1 + \|y_{k+0.5} - x^*\|\right)}{\sqrt{S_k}} \\
=\ & D_{k,p}(x^*)\|x^* - y_{k+0.5}\| + D_{k,p}(x^*)\|x^* - y_{k+0.5}\|^2.
\end{aligned} \tag{4.1.39}$$

Taking $(\mathbb{E}[|\cdot|^q|\mathcal{F}_k])^{\frac{1}{q}}$ in (4.1.39) and using $\left(\mathbb{E}[|\cdot|^q|\hat{\mathcal{F}}_k]\right)^{\frac{1}{q}} \leq \left(\mathbb{E}[|\cdot|^p|\hat{\mathcal{F}}_k]\right)^{\frac{1}{p}}$, $\left(\mathbb{E}\left[\left(\mathbb{E}[|\cdot|^q|\hat{\mathcal{F}}_k]\right)|\mathcal{F}_k\right]\right)^{\frac{1}{q}} = (\mathbb{E}[|\cdot|^q|\mathcal{F}_k])^{\frac{1}{q}}$, we deduce that

$$\begin{aligned}
&\left(\mathbb{E}\left[\left(\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle\right|^p | \hat{\mathcal{F}}_k\right]\right)^{\frac{q}{p}} | \mathcal{F}_k\right]\right)^{\frac{1}{q}} \\
\geq & \left(\mathbb{E}\left[\left(\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle\right|^q | \hat{\mathcal{F}}_k\right]\right)|\mathcal{F}_k\right]\right)^{\frac{1}{q}} \\
= & \left(\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle\right|^q | \mathcal{F}_k\right]\right)^{\frac{1}{q}}.
\end{aligned} \tag{4.1.40}$$

Once again, using the properties $\left(\mathbb{E}\left[\left(\mathbb{E}[|\cdot|^q|\hat{\mathcal{F}}_k]\right)|\mathcal{F}_k\right]\right)^{\frac{1}{q}} = (\mathbb{E}[|\cdot|^q|\mathcal{F}_k])^{\frac{1}{q}}$, $\left(\mathbb{E}[|\cdot|^q|\hat{\mathcal{F}}_k]\right)^{\frac{1}{q}} \leq \left(\mathbb{E}[|\cdot|^p|\hat{\mathcal{F}}_k]\right)^{\frac{1}{p}}$, and combining them with (4.1.36), (4.1.37), (4.1.39) (4.1.40) and Minkowski's inequality, we deduce that

$$\begin{aligned}
&\left(\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle\right|^q | \mathcal{F}_k\right]\right)^{\frac{1}{q}} \\
\leq & \left(\mathbb{E}\left[\left(\mathbb{E}\left[\left|\langle x^* - y_{k+0.5}, t_k \hat{\epsilon}_2^k \rangle\right|^p | \hat{\mathcal{F}}_k\right]\right)^{\frac{q}{p}} | \mathcal{F}_k\right]\right)^{\frac{1}{q}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\mathbb{E}\left[\left(D_{k,p}(x^*)\|x^* - y_{k+0.5}\| + D_{k,p}(x^*)\|x^* - y_{k+0.5}\|^2\right)^q |\mathcal{F}_k\right]\right)^{\frac{1}{q}} \\
&\leq D_{k,p}(x^*)\left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^q|\mathcal{F}_k\right]\right)^{\frac{1}{q}} + D_{k,p}(x^*)\left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^{2q}|\mathcal{F}_k\right]\right)^{\frac{1}{q}} \\
&\leq D_{k,p}(x^*)\left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^p|\mathcal{F}_k\right]\right)^{\frac{1}{p}} + D_{k,p}(x^*)\left(\mathbb{E}\left[\|x^* - y_{k+0.5}\|^p|\mathcal{F}_k\right]\right)^{\frac{2}{p}} \\
&\leq D_{k,p}(x^*)\left(I_{k,p}(x^*)\|x_k - x^*\| + D_{k,p}(x^*)\right) \\
&\quad + D_{k,p}(x^*)\left(2I_{k,p}(x^*)^2\|x_k - x^*\|^2 + 2D_{k,p}(x^*)^2\right) \\
&\leq D_{k,p}(x^*)\left(\frac{1}{2}I_{k,p}(x^*)^2 + \frac{1}{2}\|x_k - x^*\|^2 + D_{k,p}(x^*)\right) \\
&\quad + D_{k,p}(x^*)\left(2I_{k,p}(x^*)^2\|x_k - x^*\|^2 + 2D_{k,p}(x^*)^2\right) \\
&= \left(\frac{1}{2}D_{k,p}(x^*) + 2D_{k,p}(x^*)I_{k,p}(x^*)^2\right)\|x_k - x^*\|^2 \\
&\quad + \frac{1}{2}D_{k,p}(x^*)I_{k,p}(x^*)^2 + D_{k,p}(x^*)^2 + 2D_{k,p}(x^*)^3.
\end{aligned}
$$

□

The bound for the second-order moment of the oracle error sequence $\{\hat{\epsilon}_3^k\}$, which is not a martingale difference sequence, can be found in [38]. We present this result as a lemma.

Lemma 4.9. *[38] Suppose that Assumption 4.2 holds. Let $\xi^N := \{\xi_j\}_{j=1}^N$ be an i.i.d. sample of ξ, and $t_N : \Xi \to [0, \gamma]$ be a random variable for some $\gamma \in (0, 1]$. Let $y(\xi^N, x, t) := P_\Omega(x - t\hat{H}(\xi^N, x))$. Then for any $p \geq 2$, $x \in \mathbb{R}^n$ and $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$, there exist positive constants $\{c_i\}_{i=1}^4$ (depending on n, p and $L_{2p}\gamma$) such that,*

$$
\left(\mathbb{E}\left[\|\hat{\epsilon}(\xi^N, y(\xi^N, x, t_N))\|^p\right]\right)^{\frac{1}{p}} \leq \frac{c_1\left(\mathbb{E}\left[\|\epsilon(\xi, x^*)\|^{2p}\right]\right)^{\frac{1}{2p}} + \hat{L}_{2p}\|x - x^*\|}{\sqrt{N}},
$$

where $\hat{L}_{2p} := c_2 L_2 + c_3 L_p + c_4 L_{2p}$, with L_p defined in Remark 4.2.

For any $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi,\cdot)])$, we define

$$D_p(x^*) := \gamma B_p \delta_p(x^*) \quad \text{and} \quad I_p(x^*) := 1 + \gamma L + D_p(x^*). \tag{4.1.41}$$

Then, combining (4.1.33) and (4.1.41), we obtain

$$D_{k,p}(x^*) = \frac{1}{\sqrt{S_k}} D_p(x^*) \quad \text{and} \quad I_{k,p}(x^*) = 1 + \gamma L + \frac{1}{\sqrt{S_k}} D_p(x^*).$$

Moreover, we define

$$\begin{aligned} P(x^*) :=& 2(1-3\mu^2)(M+1)D_2(x^*)^2 + 6(M+1)D_2(x^*)^2(1+2D_2(x^*)^2) \\ &+ 6\gamma^2(M+1)c_1^2 \left(\mathbb{E}\left[\|\epsilon(\xi,x^*)\|^4\right]\right)^{\frac{1}{2}} \\ &+ 2(M+1)\left(\frac{1}{2}D_2(x^*)I_2(x^*)^2 + D_2(x^*)^2 + 2D_2(x^*)^3\right) \end{aligned} \tag{4.1.42}$$

and

$$\begin{aligned} Q(x^*) :=& 2(1-3\mu^2)(M+1)D_2(x^*)^2 + 12(M+1)D_2(x^*)^2 I_2(x^*)^2 \\ &+ 6\gamma^2(M+1)\hat{L}_4 + 2(M+1)\left(\frac{1}{2}D_2(x^*) + 2D_2(x^*)I_2(x^*)^2\right). \end{aligned} \tag{4.1.43}$$

Lemma 4.10. *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Let $\{x_k\}$ be the sequence generated by the stochastic EG-Anderson(1) algorithm, and define $\iota := \frac{(\min\{\rho\mu,\gamma\})^2}{|L(\xi)|_2^2}$. Then, there exists $\hat{\beta} \in (0, \frac{1}{2}]$ such that for any $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi,\cdot)])$,*

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_k] \leq& \left(1 + \frac{1}{\sqrt{S_k}} Q(x^*)\right) \|x_k - x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta}\iota}{2\gamma^2} r_\gamma(x_k)^2 \\ &+ \frac{1}{\sqrt{S_k}} P(x^*) + \varrho_k, \end{aligned}$$

where ϱ_k is as defined in Lemma 4.5.

Proof Taking $\mathbb{E}[\cdot|\mathcal{F}_k]$ in (4.1.26) and using the fact that $x_k \in \mathcal{F}_k$, we obtain

$$\begin{aligned}\mathbb{E}[\|x_{k+1}-x^*\|^2|\mathcal{F}_k] \leq & \|x_k-x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta}}{2\gamma^2} r_\gamma(x_k)^2 \mathbb{E}[t_k^2|\mathcal{F}_k] \\ & + (1-3\mu^2)(M+1)\gamma^2 \mathbb{E}[\|\hat{\epsilon}_1^k\|^2|\mathcal{F}_k] \\ & + 3\gamma^2(M+1)\mathbb{E}[\|\hat{\epsilon}_2^k\|^2|\mathcal{F}_k] + 3\gamma^2(M+1)\mathbb{E}[\|\hat{\epsilon}_3^k\|^2|\mathcal{F}_k] \\ & + 2(M+1)\mathbb{E}\left[\left|\langle x^*-y_{k+0.5}, t_k\hat{\epsilon}_2^k\rangle\right| |\mathcal{F}_k\right] + \varrho_k.\end{aligned} \tag{4.1.44}$$

In the following, we will bound the terms in (4.1.44) separately.

(i) From Lemma 4.8, we have

$$\mathbb{E}[\|\hat{\epsilon}_1^k\|^2|\mathcal{F}_k] \leq 2B_2^2\delta_2(x^*)^2\frac{(1+\|x_k-x^*\|^2)}{S_k}, \tag{4.1.45}$$

$$\mathbb{E}[\|\hat{\epsilon}_2^k\|^2|\mathcal{F}_k] \leq \frac{2}{S_k}B_2^2\delta_2(x^*)^2\left(1+2I_{k,2}(x^*)^2\|x_k-x^*\|^2+2D_{k,2}(x^*)^2\right) \tag{4.1.46}$$

and

$$\begin{aligned}& \mathbb{E}\left[\left|\langle x^*-y_{k+0.5}, t_k\hat{\epsilon}_2^k\rangle\right| |\mathcal{F}_k\right] \\ \leq & \left(\frac{1}{2}D_{k,2}(x^*)+2D_{k,2}(x^*)I_{k,2}(x^*)^2\right)\|x_k-x^*\|^2 \\ & + \frac{1}{2}D_{k,2}(x^*)I_{k,2}(x^*)^2 + D_{k,2}(x^*)^2 + 2D_{k,2}(x^*)^3.\end{aligned} \tag{4.1.47}$$

(ii) From Lemma 4.9 and the fact that ξ^k is independent of $\mathcal{F}_k$, it follows that

$$\begin{aligned}\mathbb{E}[\|\hat{\epsilon}_3^k\|^2|\mathcal{F}_k] = & \mathbb{E}[\|\hat{\epsilon}(\xi^k, y(\xi^k, x_k, t_k))\|^2] \\ \leq & \frac{2c_1^2\left(\mathbb{E}\left[\|\epsilon(\xi,x^*)\|^4\right]\right)^{\frac{1}{2}} + 2\hat{L}_4^2\|x_k-x^*\|^2}{S_k}.\end{aligned} \tag{4.1.48}$$

From Lemma 4.3, we get

$$\mathbb{E}[t_k^2|\mathcal{F}_k] \geq \frac{\left(\min\{\rho\mu, \gamma\}\right)^2}{\mathbb{E}[L(\xi)^2]}. \tag{4.1.49}$$

Substituting (4.1.45)-(4.1.49) into (4.1.44), we conclude that

$$\begin{aligned}\mathbb{E}[\|x_{k+1}-x^*\|^2|\mathcal{F}_k] \leq \left(1+\frac{1}{\sqrt{S_k}}Q(x^*)\right)\|x_k-x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta}\iota}{2\gamma^2}r_\gamma(x_k)^2 \\ + \frac{1}{\sqrt{S_k}}P(x^*)+\varrho_k.\end{aligned} \tag{4.1.50}$$

□

Based on the above lemmas, we present the main conclusion.

Theorem 4.2. *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Let $\{x_k\}$ be the sequence generated by the stochastic EG-Anderson(1) algorithm. Then, a.s. the sequence $\{x_k\}$ converges to a solution of the problem* (4.1.1) *and $r_\gamma(x_k)$ converges to 0.*

Proof Let x^* be a solution of the problem (4.1.1). From Lemma 4.10, we find

$$\begin{aligned}\mathbb{E}[\|x_{k+1}-x^*\|^2|\mathcal{F}_k] \leq \left(1+\frac{1}{\sqrt{S_k}}Q(x^*)\right)\|x_k-x^*\|^2 - \frac{(1-3\mu^2)\hat{\beta}\iota}{2\gamma^2}r_\gamma(x_k)^2 \\ + \frac{1}{\sqrt{S_k}}P(x^*)+\varrho_k.\end{aligned} \tag{4.1.51}$$

Since $\sum_{k=0}^{\infty}\frac{1}{\sqrt{S_k}}<\infty$ (Assumption 4.3) and $\sum_{k=0}^{\infty}\varrho_k<\infty$ (Lemma 4.5), it follows that $\sum_{k=0}^{\infty}\frac{1}{\sqrt{S_k}}Q(x^*)<\infty$ and $\sum_{k=0}^{\infty}\left(\frac{1}{\sqrt{S_k}}P(x^*)+\varrho_k\right)<\infty$. Using $\mu\in(0,\frac{\sqrt{3}}{3})$ and applying Theorem 4.1 with $y_k=\|x_k-x^*\|^2$, $a_k=\frac{1}{\sqrt{S_k}}Q(x^*)$, $u_k=\frac{(1-3\mu^2)\hat{\beta}\iota}{2\gamma^2}r_\gamma(x_k)^2$ and $b_k=\frac{1}{\sqrt{S_k}}P(x^*)+\varrho_k$, we conclude that a.s. $\|x_k-x^*\|^2$ converges and $\sum_{k=0}^{\infty}r_\gamma(x_k)^2<\infty$.

In particular, a.s. $\{x_k\}$ is bounded, which implies that, a.s., there exists a cluster point $\hat{x}$ and a subsequence $\{x_{s_j}\}$ satisfying

$$\lim_{j\to\infty}x_{s_j}=\hat{x}. \tag{4.1.52}$$

Moreover, we have that a.s.

$$\lim_{k\to\infty} r_\gamma(x_k)^2 = \lim_{k\to\infty} \|x_k - P_\Omega\left(x_k - \gamma\mathbb{E}[T(\xi, x_k)]\right)\|^2 = 0. \tag{4.1.53}$$

From (4.1.53) and the continuity of both H and P_Ω, we deduce that a.s. $\|\hat{x} - P_\Omega(\hat{x} - \gamma\mathbb{E}[T(\xi, \hat{x})])\| = 0$. Together with Lemma 4.1, this implies that a.s. $\hat{x} \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$. Since a.s. $\|x_k - x^*\|$ converges for any $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$, it follows that a.s. $\|x_k - \hat{x}\|$ also converges. Together with (4.1.52), we have that a.s.

$$\lim_{k\to\infty} \|x_k - \hat{x}\| = \lim_{j\to\infty} \|x_{s_j} - \hat{x}\| = 0.$$

□

Before presenting the conclusion about the convergence rate, we first provide an upper bound for $\sup_{k\geq k_0} \mathbb{E}\left[\|x_k - x^*\|^2\right]$.

Lemma 4.11. *Suppose Assumptions 4.1, 4.2 and 4.3 hold. Let $\{x_k\}$ be the sequence generated by the stochastic EG-Anderson(1) algorithm. Let $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$, and choose $k_0 \in \mathbb{N}$, $\phi \in (0, 1)$ and $\psi > 0$ such that $\sum_{k=k_0}^{\infty} \frac{1}{\sqrt{S_k}} < \frac{\phi}{Q(x^*)}$ and $\sum_{k=k_0}^{\infty} \varrho_k < \psi$. Then, we have*

$$\sup_{k\geq k_0} \mathbb{E}[\|x_k - x^*\|^2] \leq \frac{\mathbb{E}[\|x_{k_0} - x^*\|^2] + P(x^*)\frac{\phi}{Q(x^*)} + \psi}{1 - \phi}.$$

Proof Taking expectation in (4.1.51) and applying $\mathbb{E}\left[\mathbb{E}\left[\cdot|\mathcal{F}_k\right]\right] = \mathbb{E}[\cdot]$, we drop the negative term to obtain

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 + \frac{1}{\sqrt{S_k}}Q(x^*)\right)\mathbb{E}\left[\|x_k - x^*\|^2\right] + \frac{1}{\sqrt{S_k}}P(x^*) + \varrho_k. \tag{4.1.54}$$

Let $d_i = \|x_i - x^*\|$ for $i \in \mathbb{N}_0$ and $k > k_0$. Then, summing (4.1.54) from k_0 to $k-1$ yields

$$\mathbb{E}[d_k^2] \leq \mathbb{E}[d_{k_0}^2] + Q(x^*)\sum_{i=k_0}^{k-1} \frac{1}{\sqrt{S_i}}\mathbb{E}[d_i^2] + P(x^*)\sum_{i=k_0}^{k-1} \frac{1}{\sqrt{S_i}} + \sum_{i=k_0}^{k-1} \varrho_i. \tag{4.1.55}$$

For any $a > 0$, we define $\varpi_a := \inf\{k > k_0 : \mathbb{E}[d_k^2]^{\frac{1}{2}} > a\}$. Assume that $\mathbb{E}[d_k^2]$ is unbounded, then $\varpi_a < \infty$ for all $a > 0$. By choosing $k = \varpi_a$ in (4.1.55) and considering that $\sum_{k=k_0}^{\infty} \frac{1}{\sqrt{S_k}} < \frac{\phi}{Q(x^*)}$ and $\sum_{k=k_0}^{\infty} \varrho_k < \psi$, we obtain

$$\begin{aligned} a^2 < \mathbb{E}[d_{\varpi_a}^2] \leq & \mathbb{E}[d_{k_0}^2] + Q(x^*) \sum_{i=k_0}^{\varpi_a - 1} \frac{1}{\sqrt{S_i}} \mathbb{E}[d_i^2] + P(x^*) \sum_{i=k_0}^{\varpi_a - 1} \frac{1}{\sqrt{S_i}} + \sum_{i=k_0}^{\varpi_a - 1} \varrho_i \\ \leq & \mathbb{E}[d_{k_0}^2] + \phi a^2 + P(x^*) \frac{\phi}{Q(x^*)} + \psi, \end{aligned}$$

which implies

$$a^2 \leq \frac{\mathbb{E}[d_{k_0}^2] + P(x^*) \frac{\phi}{Q(x^*)} + \psi}{1 - \phi}.$$

This contradicts the arbitrariness of a, thus $\mathbb{E}[d_k^2]$ is bounded and

$$\sup_{k \geq k_0} \mathbb{E}[d_k^2] \leq \frac{\mathbb{E}[d_{k_0}^2] + P(x^*) \frac{\phi}{Q(x^*)} + \psi}{1 - \phi}.$$

□

Remark 4.4. *Let $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$. From Lemma 4.11, it follows that under Assumptions 4.1, 4.2 and 4.3, there exist $k_0 \in \mathbb{N}$, $\phi \in (0, 1)$ and $\psi > 0$ such that*

$$\sup_{k \geq 0} \mathbb{E}[\|x_k - x^*\|^2] \leq \frac{\max_{k=0,\cdots,k_0} \mathbb{E}[\|x_{k_0} - x^*\|^2] + P(x^*) \frac{\phi}{Q(x^*)} + \psi}{1 - \phi}. \tag{4.1.56}$$

Theorem 4.3. *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Let $\{x_k\}$ be the sequence generated by the stochastic EG-Anderson(1) algorithm. Then we have*

$$\min_{0 \leq k \leq N} \mathbb{E}[r_\gamma(x_k)^2] = O\left(\frac{1}{N}\right).$$

Proof Let $x^* \in \mathrm{SSOL}(\Omega, \mathbb{E}[T(\xi, \cdot)])$. Taking expectation in (4.1.51) and using $\mathbb{E}\left[\mathbb{E}\left[\cdot | \mathcal{F}_k\right]\right] = \mathbb{E}[\cdot]$, we recursively sum the resulting inequality from $k = 0$ to $k = N$

to obtain

$$\frac{(1-3\mu^2)\hat{\beta}\iota}{2\gamma^2}\sum_{k=0}^{N}\mathbb{E}\left[r_\gamma(x_k)^2\right] \leq \|x_0-x^*\|^2 + Q(x^*)\sum_{k=0}^{N}\frac{1}{\sqrt{S_k}}\mathbb{E}\left[\|x_k-x^*\|^2\right] + P(x^*)\sum_{k=0}^{N}\frac{1}{\sqrt{S_k}} + \sum_{k=0}^{N}\varrho_k.$$

Since $\sum_{k=0}^{\infty}\frac{1}{\sqrt{S_k}} < \infty$ (Assumption 4.3), $\sum_{k=0}^{\infty}\varrho_k < \infty$ (Lemma 4.5) and (4.1.56), we know that there exists $\Phi > 0$ such that

$$\sum_{k=0}^{N}\mathbb{E}\left[r_\gamma(x_k)^2\right] \leq \frac{2\gamma^2}{(1-3\mu^2)\hat{\beta}\iota}\Phi,$$

which implies

$$\min_{0\leq k\leq N}\mathbb{E}\left[r_\gamma(x_k)^2\right] \leq \frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[r_\gamma(x_k)^2\right] \leq \frac{1}{N+1}\frac{2\gamma^2}{(1-3\mu^2)\hat{\beta}\iota}\Phi.$$

□

Theorem 4.4. *Suppose that Assumptions 4.1, 4.2 and 4.3 hold. Let $\{x_k\}$ be the sequence generated by the stochastic EG-Anderson(1) algorithm. Set $S_k := \mathcal{N}\lceil(k+\lambda)^2(\ln(k+\lambda))^{2+2b}\rceil$ for $\mathcal{N} = \mathcal{O}(n^2), b > 0$ and $\lambda > 2$. Given $\varsigma > 0$, the stochastic EG-Anderson(1) algorithm achieves the tolerance*

$$\min_{0\leq k\leq N}\mathbb{E}\left[r_\gamma(x_k)^2\right] \leq \varsigma$$

after $N = \mathcal{O}(b^{-1}\varsigma^{-1})$ iterations and the oracle complexity $\sum_{k=0}^{N}(1+m_k)S_k$ is bounded above by

$$b^{-3}\log_{\frac{1}{\rho}}\left(\frac{\gamma\hat{L}(\xi^k)}{\min\{\rho\mu,\gamma\}}\right)\ln(b^{-1}\varsigma^{-1})^{2+2b}\mathcal{O}(n^2\varsigma^{-3})$$

with probability 1, where m_k is the number of oracle calls used in the line search scheme (4.1.6) *at the k-iteration and $\hat{L}(\xi^k)$ is defined as in Lemma 4.3.*

Proof From the form of S_k, we can compute that

$$\sum_{k=0}^{\infty} \frac{1}{\sqrt{S_k}} \leq \int_{-1}^{\infty} \frac{dt}{\sqrt{\mathcal{N}}(t+\lambda)\left(\ln(t+\lambda)\right)^{1+b}} = \frac{1}{\sqrt{\mathcal{N}}b\left(\ln(\lambda-1)\right)^{b}} < \infty. \tag{4.1.57}$$

According to the definitions of $\{c_i\}_{i=1}^4$, $\hat{L}_{2p}$ in Lemma 4.9, $P(x^*)$ in (4.1.42) and $Q(x^*)$ in (4.1.43) (which depend on n), along with Theorem 4.3, there exists a constant $U > 0$ such that for any $k \in \mathbb{N}_0$, $\min_{0\leq k\leq N} \mathbb{E}\left[r_\gamma(x_k)^2\right] \leq nU(\sqrt{\mathcal{N}}bN)^{-1}$. Hence, given $\varsigma > 0$, we obtain

$$\min_{0\leq k\leq N} \mathbb{E}\left[r_\gamma(x_k)^2\right] \ \leq \varsigma$$

after $N = \mathcal{O}(n(\sqrt{\mathcal{N}}b\varsigma)^{-1})$ iterations.

Let m_k be the number of line search in the iteration k. After N iterations, the oracle complexity is upper bounded by

$$\begin{aligned}
&\sum_{k=0}^{N}(1+m_k)S_k \\
&\leq \left(1+\max_{k\in\{0,\cdots,N\}} m_k\right)\sum_{k=0}^{N}\left(\mathcal{N}(k+\lambda)^2(\ln(k+\lambda))^{2+2b}+1\right) \\
&= \left(1+\max_{k\in\{0,\cdots,N\}} m_k\right)\sum_{k=\lambda}^{N+\lambda}\left(\mathcal{N}k^2(\ln k)^{2+2b}+1\right) \\
&\leq \left(1+\max_{k\in\{0,\cdots,N\}} m_k\right)\left(\mathcal{N}(N+1)(N+\lambda)^2(\ln(N+\lambda))^{2+2b}+1\right) \\
&\leq \left(1+\max_{k\in\{0,\cdots,N\}} m_k\right)\left(\mathcal{N}(b^{-1}\varsigma^{-1}+1)(b^{-1}\varsigma^{-1}+\lambda)^2(\ln(b^{-1}\varsigma^{-1}+\lambda))^{2+2b}+1\right).
\end{aligned} \tag{4.1.58}$$

Moreover, Lemma 4.3 implies that a.s. $m_k \leq \log_{\frac{1}{\rho}}\left(\frac{\gamma\hat{L}(\xi^k)}{\min\{\rho\mu,\gamma\}}\right)$. Combining this with (4.1.58) and $\mathcal{N} = \mathcal{O}(n^2)$ yields the claimed bound on $\sum_{k=0}^{N}(1+m_k)S_k$. □

4.2 Numerical experiments

In this section, we conduct numerical examples to compare the stochastic EG-Anderson(1) algorithm with the stochastic EG algorithm that incorporates line search (4.1.6). In addition, we also perform numerical experiments on portfolio selection problems using real data. All the codes are written in Matlab (R2023b) and run on a MacBook Air (16.00GB of RAM).

4.2.1 Simple examples

In the following experiments of this subsection, the stopping criterion for the two examples is set as

$$r(\xi^k, x_k) = \left\|F_1(\xi^k, x_k)\right\| = \left\|P_\Omega\left(x_k - \hat{H}(\xi^k, x_k)\right) - x_k\right\| < 10^{-2}$$

or when the number of iterations exceeds 200. The parameters for the stochastic EG-Anderson(1) algorithm are specified as follows:

$$\nu = 30, M = 5000, \tau = 0.6, \rho = 0.8, \mu = 0.5. \tag{4.2.1}$$

In the figures and tables presented in this subsection, 'Sec.' denotes the CPU time measured in seconds, while 'Iter.' indicates the total number of iterations. The numbers in parentheses denote the number of executed Anderson steps (4.1.8). Additionally, the algorithm that performs best regarding the average number of iterations and CPU time is emphasized in bold for each combination of dimension n and parameter γ.

Example 4.1. *[13] Consider the stochastic complementarity problem*

$$\mathbb{E}[T(\xi, x)] \geq \mathbf{0}, \quad x \geq \mathbf{0}, \quad x^{\mathrm{T}}\mathbb{E}[T(\xi, x)] = 0, \tag{4.2.2}$$

where $T(\xi, x) := D(\xi, x) + Q(\xi)x + q(\xi)$ *with* $D(\xi, x)$ *and* $Q(\xi)x + q(\xi)$ *to be its nonlinear and linear parts, respectively. Assume that*

- $D_i(\xi, x) := d_i(\xi) \cdot \arctan(a_i(\xi)x_i)$, *where* $d(\xi)$ *and* $a(\xi)$ *are vectors whose elements obey uniform distributions on (0,1);*

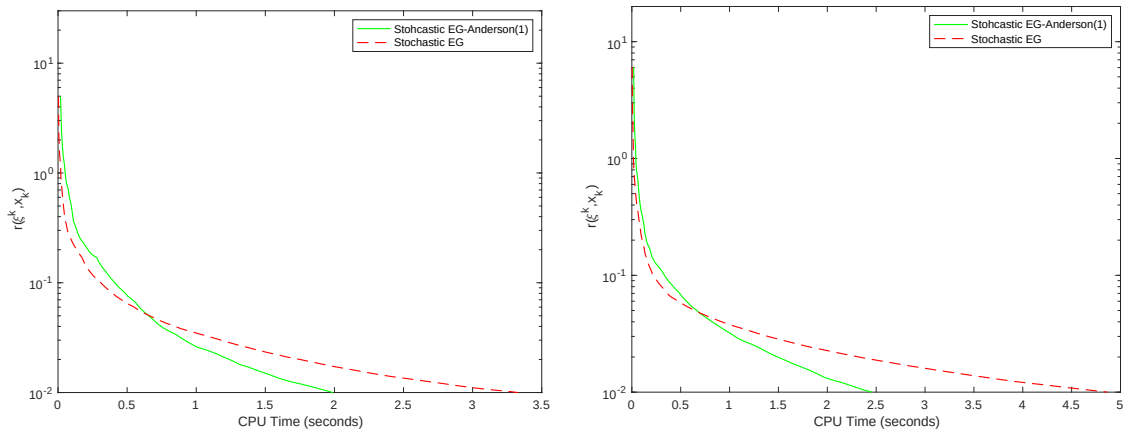
- $Q(\xi) := B + Y(\xi)$, *where* $B \in \mathbb{R}^{n\times n}$ *is a skew symmetric matrix whose elements obey uniform distributions on (0,5) and* $Y(\xi) \in \mathbb{R}^{n\times n}$ *is a diagonal matrix whose elements obey uniform distributions on (0,2);*

- $q(\xi) \in \mathbb{R}^n$ *is a vector whose elements obey uniform distributions on (-3,3).*

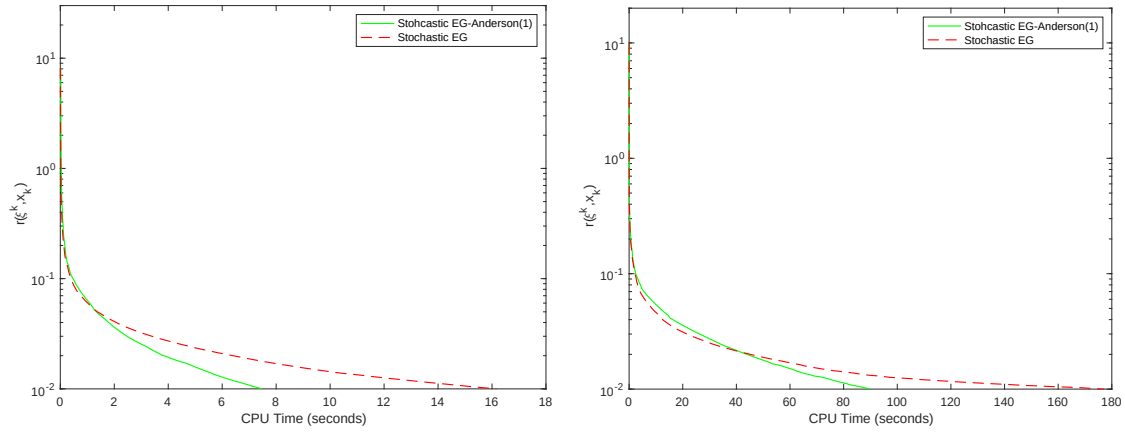
From the above generation method, it can be seen that $\mathbb{E}[T(\xi, \cdot)]$ is monotone. Moreover, the stochastic complementarity problem given in (4.2.2) can be reformulated as SVI$(\Omega, \mathbb{E}[T(\xi, \cdot)])$ by choosing $\Omega = \{x : x \geq \mathbf{0}\}$.

Let $S_k := \mathcal{N}\lceil(k + \lambda)^2(\ln(k + \lambda))^{2+2b}\rceil$ with $\lambda = 2 + 10^{-4}$, $b = 10^{-4}$ and $\mathcal{N} = 2$. Using the stochastic EG-Anderson(1) and the stochastic EG algorithms to solve Example 4.1, the experimental results are presented in Figure 4.1. This figure depicts the variation of $r(\xi^k, x_k)$ with respect to CPU time for dimensions $n = 50, 70, 100,$ and 150. It can be seen that, upon reaching the stopping criterion, the stochastic EG-Anderson(1) algorithm requires less CPU time than that of the stochastic EG algorithm.

We also conducted experiments with S_k set to $\mathcal{N}\lceil(k+\lambda)(\ln(k+\lambda))^{1+b}\rceil$, where $\lambda = 5$, $b = 0.1$ and $\mathcal{N} = 20$. The corresponding results are shown in Table 4.1. Table 4.1 presents the average number of iterations and average CPU time for two algorithms across different dimensions based on ten experiments. As can be seen from Table 4.1, even when the value of S_k is smaller than the requirement in Assumption 4.3, the stochastic EG-Anderson(1) algorithm can still successfully solve this problem. Moreover, in most cases, its numerical performance, in terms of both the number of iterations and CPU time, is better than that of the stochastic EG algorithm.



(a) $n = 50$, $\gamma = 0.05$ (b) $n = 70$, $\gamma = 0.03$



(c) $n = 100$, $\gamma = 0.02$ (d) $n = 150$, $\gamma = 0.01$

Figure 4.1: Comparisons of the convergence behaviours of the stochastic EG-Anderson(1) and the stochastic EG algorithms for Example 4.1

Example 4.2. *[42] Consider the stochastic fractional convex quadratic problem*

$$\min_{x\in\Omega} \mathbb{E}\left[\frac{f(\xi,x)}{g(x)}\right], \tag{4.2.3}$$

where $f(\xi,x) := 0.5x^{\mathrm{T}}\left(0.025U^{\mathrm{T}}U + 0.025\frac{\|U^{\mathrm{T}}U\|_F}{\|V(\xi)\|_F}V(\xi)\right)x + 0.5\left((c+\bar{c}(\xi))^{\mathrm{T}}x + 4n\right)^2$, $g(x) := r^{\mathrm{T}}x + d + 4n$, *and* $\Omega := \{x : \mathbf{0} \le x \le 4\mathbf{e}\}$.

We note that $V(\xi)$ and $\bar{c}(\xi)$ are randomly generated, with $V(\xi)$ drawn from a standard normal distribution and $\bar{c}(\xi)$ drawn from a uniform distribution on [0, 1],

Sec.(avr) Iter.(avr)	stochastic EG-Anderson(1)	stochastic EG
$n = 10$ $\gamma = 0.1$	0.0407 **36.7** (27.7)	**0.0355** 40
$n = 20$ $\gamma = 0.08$	**0.1376** **62.3** (41.9)	0.1390 67.5
$n = 30$ $\gamma = 0.03$	**0.1258** **59.1** (46.1)	0.1372 67.5
$n = 50$ $\gamma = 0.02$	**0.3442** **80** (54.1)	0.4192 94
$n = 100$ $\gamma = 0.02$	**0.7064** **94.9** (63.8)	0.8094 106.3

Table 4.1: Comparisons of the two algorithms for Example 4.1

respectively. U and c are deterministic constants generated once from the standard normal distribution, while r and d are generated once from uniform distributions on [0,5].

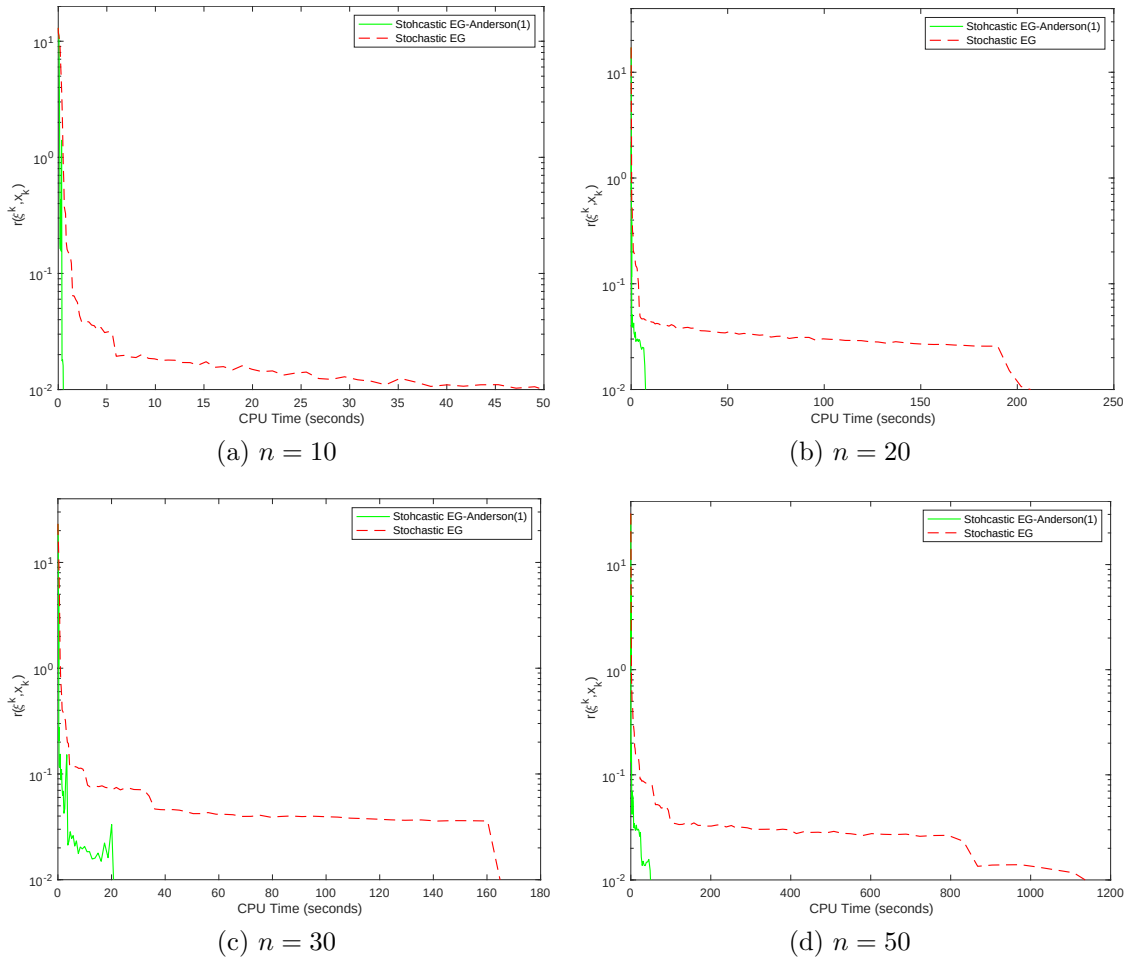
It is easy to see that problem (4.2.3) is equivalent to the SVI($\Omega, \mathbb{E}[T(\xi, \cdot)]$), where

$$\mathbb{E}[T(\xi, x)] = \nabla_x \mathbb{E}\left[\frac{f(\xi, x)}{g(x)}\right] = \mathbb{E}\left[\frac{\nabla_x f(\xi, x) \cdot g(x) - f(\xi, x) \cdot \nabla g(x)}{[g(x)]^2}\right]. \tag{4.2.4}$$

By [42], we know that $\mathbb{E}[T(\xi, \cdot)]$ is pseudomonotone on Ω.

Let $\gamma = 0.9$ and $S_k := \mathcal{N}\lceil(k+\lambda)^2(\ln(k+\lambda))^{2+2b}\rceil$ with $\lambda = 2+10^{-4}$, $b = 10^{-4}$ and $\mathcal{N} = 2$. Starting from the same random initial point, Figure 4.2 presents the variation of $r(\xi^k, x_k)$ with respect to CPU time for different dimensions. It can be observed that the numerical performance of the stochastic EG-Anderson (1) algorithm is superior to that of the stochastic EG algorithm.

We also performed experiments where S_k was set to $\mathcal{N}\lceil(k+\lambda)(\ln(k+\lambda))^{1+b}\rceil$ with $\lambda = 5$, $b = 0.1$ and $\mathcal{N} = 20$. Table 4.2 reports the average number of iterations and the average CPU time for both algorithms across various dimensions, based on ten experimental runs. As indicated in Table 4.2, the stochastic EG-Anderson(1)



(a) $n = 10$ (b) $n = 20$

(c) $n = 30$ (d) $n = 50$

Figure 4.2: Comparisons of the convergence behaviours of the stochastic EG-Anderson(1) and the stochastic EG algorithms for Example 4.2

algorithm is able to successfully solve this problem even when S_k is chosen to be smaller than the value specified in Assumption 4.3. Furthermore, the results show that, across all tested dimensions, the stochastic EG-Anderson(1) algorithm consistently requires fewer iterations and less CPU time than the stochastic EG algorithm to meet the stopping criteria.

Sec.(avr) Iter.(avr)	stochastic EG-Anderson(1)	stochastic EG
$n = 10$	**0.2930** **24.9** (16.6)	1.5853 62.4
$n = 20$	**1.6086** **44.3** (25.9)	4.6915 77
$n = 30$	**3.3112** **49.1** (29.6)	7.4201 75.7
$n = 50$	**14.4326** **69.9** (39)	42.2535 119.7
$n = 100$	**113.4942** **96.5** (53.9)	206.2057 124

Table 4.2: Comparisons of the two algorithms for Example 4.2

4.2.2 Portfolio management

Example 4.3. *The Markowitz mean-variance model [52] for portfolio selection can be formulated as*

$$\begin{aligned} &\min_{w\in\mathbb{R}^n} \frac{1}{2} w^{\mathrm{T}}\Sigma w - \varrho^{\mathrm{T}} w \\ &\text{s.t.} \;\; \mathbf{0} \le w \le a, \;\; \mathbf{e}^{\mathrm{T}} w = 1, \end{aligned} \tag{4.2.5}$$

where $w \in \mathbb{R}^n$ denotes the weights of the assets in the portfolio, $\varrho \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$ are the mean vector and the covariance matrix of returns on assets, respectively. $a \in \mathbb{R}^n_+$ is the upper bound enforced on w.

In the following experiments, we set $a = 0.2\mathbf{e}$. It is clear that this is equivalent to the linear complementarity problem defined as $\mathrm{LCP}(\hat{q}, \hat{M})$, which is given by:

$$\hat{M}x + \hat{q} \ge \mathbf{0}, \quad x \ge \mathbf{0}, \quad x^{\mathrm{T}}(\hat{M}x + \hat{q}) = 0, \tag{4.2.6}$$

where

$$\hat{M} = \begin{bmatrix} \Sigma & -C^{\mathrm{T}} \\ C & O \end{bmatrix}, C = \begin{bmatrix} \mathbf{e}^{\mathrm{T}} \\ -\mathbf{e}^{\mathrm{T}} \\ -I \end{bmatrix}, \hat{q} = \begin{bmatrix} -\varrho \\ -1 \\ 1 \\ a \end{bmatrix}, x = \begin{bmatrix} w \\ y \end{bmatrix},$$

with O denoting the zero matrix. The dual variables y are constrained to be in $\mathbb{R}^{n+2}_{+}$. Moreover, the LCP$(\hat{q}, \hat{M})$ in (4.2.6) is equivalent to the SVI$(\Omega, \mathbb{E}[T(\xi, \cdot)])$, where $\mathbb{E}[T(\xi, x)] = \hat{M}x + \hat{q}$ and $\Omega = \{x \in \mathbb{R}^n : x \geq \mathbf{0}\}$. It can be verified that both Assumptions 4.1 and 4.2 hold.

The stock data sets utilized in our experiments are drawn from two sources: the standard data sets available in the OR-library (referred to as Data 1-2) and those from the China A-share market (referred to as Data 3-6). For the standard data sets, we utilize weekly stock prices from 1992 to 1997 for the DAX 100 Index (Germany) and the Nikkei Index (Japan). For Data sets 3 and 4, we employ daily closing prices from the China A-share market covering the period from 2010 to 2017. Additionally, for Data sets 5 and 6, we use daily closing prices for CSI300 index stocks from the China A-share market, spanning from 2015 to 2022.

For the A-share and CSI300 index data, we remove stocks with more than 10% zero daily returns, resulting in 1,071 and 189 stocks as Data 3 and Data 5, respectively. We then rank these stocks by their average daily returns and select the top 150 and 100 stocks as Data 4 and Data 6, respectively. A description of all the data sets can be found in Table 4.3.

For each data set, let the stock closing prices be represented as $\{P_{j,i} : j \in [T], i \in [n]\}$. The n represents the number of component assets, while the T indicates the number of observations for those assets. We compute the returns for each stock using

Data sets	n	**Source**	**Index**	J	**Description**
Data 1	85	Germany	DAX 100	291	weekly prices from 1992 to 1997
Data 2	225	Japan	Nikkei 225	291	weekly prices from 1992 to 1997
Data 3	1071	China	A-share	1940	daily prices from 2010 to 2017
Data 4	150	China	A-share	1940	daily prices from 2010 to 2017
Data 5	189	China	CSI300	1986	daily prices from 2015 to 2022
Data 6	100	China	CSI300	1986	daily prices from 2015 to 2022

Table 4.3: Description of the six real data sets

the formula

$$r_{j,i} = \log \frac{P_{j+1,i}}{P_{j,i}}$$

for $j \in [T-1], i \in [n]$. This results in a return matrix $R \in \mathbb{R}^{(T-1)\times n}$ containing the elements $r_{j,i}$. Each data set is divided into two subsets in a 9:1 ratio, referred to as the training set and the testing set. The training set, known as the in-sample set, consists of the first 9/10 of the data and is used to calculate the optimal portfolio. The testing set, referred to as the out-of-sample set, includes the remaining data and is utilized to evaluate the performance of the resulting optimal portfolio.

Let $R \in \mathbb{R}^{(T-1)\times n}$ represent the return matrix of the assets over $[1, T-1]$. Based on the splitting ratio, we define $R_{in} \in \mathbb{R}^{T_{in}\times n}$ and $R_{out} \in \mathbb{R}^{T_{out}\times n}$, where $T_{in} = 2\lceil 0.9(T-1)/2 - 1\rceil$ and $T_{out} = T - 1 - T_{in}$. Specifically, R_{in} consists of the first T_{in} rows of R, while R_{out} comprises the remaining rows.

We use the stochastic EG-Anderson(1) algorithm to solve the problem (4.2.6) on the training set and find the optimal parameters w for portfolio selection. We then compare its performance on the test set with that of the Naive portfolio (the naive evenly weighted portfolio, with all weights being $\frac{1}{n}$). In practical applications, the out-of-sample portfolio performance is of greater interest, as it better reflects the portfolio's potential future returns. To enable comparison, we introduce two metrics

to assess out-of-sample performance as follows:

- the Sharpe ratio (SR) of the portfolio w, defined as

$$\mathrm{SR}(w) = \frac{\varrho_{out}^{\mathrm{T}} w}{\sqrt{w^{\mathrm{T}} \Sigma_{out} w}},$$

  where ϱ_{out} is the mean of each column in R_{out}, and Σ_{out} is the covariance matrix centered around ϱ_{out}.

- the annualized return (AR) of the portfolio w, defined as

$$\mathrm{AR}(w) = \frac{\mathrm{CR}(w)}{T_w} \times 50 \quad \text{or} \quad \frac{\mathrm{CR}(w)}{T_d} \times 250,$$

  where $\mathrm{CR}(w)$ represents the out-of-sample cumulative return for w, T_w and T_d refer to the number of weeks and days, respectively, in the out-of-sample dataset.

In general, if a strategy has higher SR and AR, it is considered superior to other strategies.

When executing the stochastic EG-Anderson(1) algorithm, the settings of some parameters in the algorithm are the same as in (4.2.1). Moreover, we set $\gamma = 0.1$ and

$$S_k := \min\{T_{in}/2, \mathcal{N}\lceil (k+\lambda)^2 (\ln(k+\lambda))^{2+2b} \rceil\}$$

with $\lambda = 2 + 10^{-4}$, $b = 10^{-4}$, and $\mathcal{N} = 2$. We generate $\varrho \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ in the following way:

$$\varrho(\xi^k) = \mathrm{mean}(R_{in}(T_{in}/2 : T_{in}/2 + S_k, :)); \quad \Sigma(\xi^k) = \mathrm{cov}(R_{in}(T_{in}/2 : T_{in}/2 + S_k, :));$$

$$\varrho(\eta^k) = \mathrm{mean}(R_{in}(1 : S_k, :)); \quad \Sigma(\eta^k) = \mathrm{cov}(R_{in}(1 : S_k, :)),$$

where $\{\xi^k\}$ and $\{\eta^k\}$ are random variables related to asset returns involved in each iteration. Then, we have

$$\hat{M}(\xi^k) = \begin{bmatrix} \Sigma(\xi^k) & -C^{\mathrm{T}} \\ C & O \end{bmatrix}, \hat{q}(\xi^k) = \begin{bmatrix} -\varrho(\xi^k) \\ -1 \\ 1 \\ a \end{bmatrix}$$

and

$$\hat{M}(\eta^k) = \begin{bmatrix} \Sigma(\eta^k) & -C^{\mathrm{T}} \\ C & O \end{bmatrix}, \hat{q}(\eta^k) = \begin{bmatrix} -\varrho(\eta^k) \\ -1 \\ 1 \\ a \end{bmatrix}.$$

The stopping criterion for this example is set as the number of iterations exceeding 2000. In the following figures and table, for simplicity, 'Naive' represents the average strategy, while 'SEGA' denotes the weights obtained by solving (4.2.6) using the stochastic EG-Anderson(1) algorithm. Table 4.4 presents the performance of the portfolio solved by the two methods on the test sets: results concerning SR and AR. According to Table 4.4, it is evident that applying the stochastic EG-Anderson(1) algorithm to solve (4.2.6) yields a portfolio strategy with higher SR and AR compared to the average strategy across all data sets. Moreover, Figure 4.3 shows the curves of SR, CR, and the objective function value (abbreviated as f) in problem (4.2.5) on the testing set as the number of iterations increases for SEGA and Naive on Data 2.

Data sets	SR		AR	
	Naive	SEGA	Naive	SEGA
Data 1	0.2701	**0.2744**	9.5974e-04	**0.0015**
Data 2	-0.2135	**0.0716**	-0.0010	**3.4974e-04**
Data 3	-0.0492	**0.0403**	-6.3285e-05	**4.9256e-05**
Data 4	0.0691	**0.1029**	8.6776e-05	**1.4832e-04**
Data 5	0.0506	**0.0534**	6.4001e-05	**8.2941e-05**
Data 6	0.0465	**0.0536**	6.5498e-05	**8.3823e-05**

Table 4.4: Performance metrics for different data sets

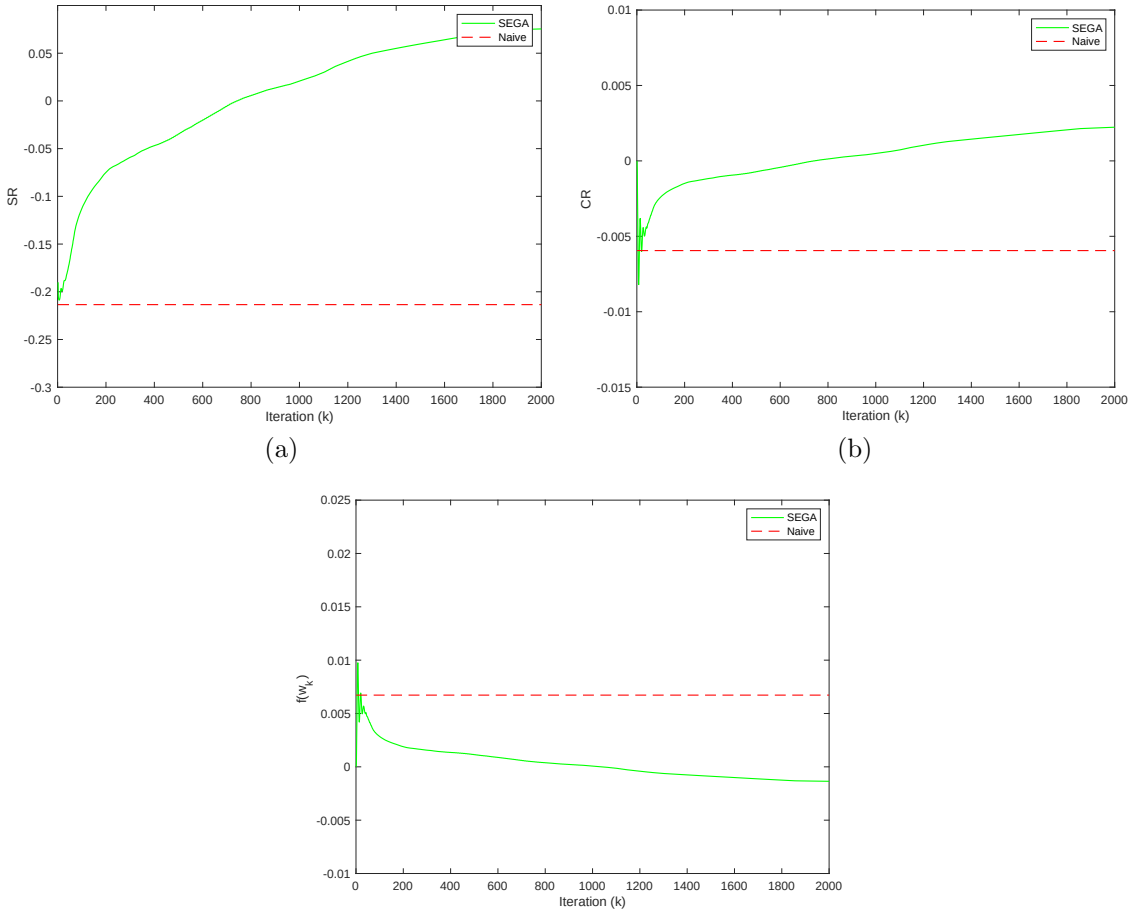


Figure 4.3: Convergence curves and comparison of SR and CR on the testing set of Data 2

Chapter 5

Conclusions and Future Work

This chapter summarizes the conclusions of this thesis and outlines potential directions for future research.

5.1 Conclusions

This thesis investigates algorithms for solving nonmonotone VIs, particularly designing accelerated algorithms for pseudomonotone VIs and SVIs. The specific conclusions are as follows.

- We propose an algorithm called EG-Anderson(1), based on the EG method and Anderson acceleration, for solving pseudomonotone VI problem (3.1.1). In this thesis, we prove that the sequence generated by the EG-Anderson(1) algorithm can converge to a solution of the problem (3.1.1) without assuming Lipschitz continuity and contraction conditions. Subsequently, when the operator is locally Lipschitz continuous, we provide the convergence rate of the EG-Anderson(1) algorithm on the residual function, which is no worse than the convergence rate of the EG method, and this condition is weaker than the global Lipschitz continuity required by the EG method. We also present an improved version of the algorithm by reducing the number of projections onto the feasible set, which is suitable for situations where projections onto the feasible

set are difficult to compute. Numerical experiments show that EG-Anderson(1) algorithm performs better than the EG method and Anderson acceleration.

- We present the stochastic EG-Anderson(1) algorithm, an extension of EG-Anderson(1) algorithm, for solving pseudomonotone SVI problem (4.1.1). Building upon the EG method and Anderson acceleration, this algorithm employs SA framework. To enhance robustness, the stochastic EG-Anderson(1) algorithm integrates a line search strategy, eliminating the need for prior knowledge of the Lipschitz constant. We demonstrate that, the sequence generated by the stochastic EG-Anderson(1) algorithm converges almost surely to a solution of the problem (4.1.1). Furthermore, we derive a sublinear convergence rate of $O(N^{-1})$ for the mean residual function and establish the algorithm's optimal oracle complexity. Numerical results indicate that the stochastic EG-Anderson(1) algorithm achieves superior performance compared to the stochastic EG method. Moreover, in the portfolio selection problem with real data, the solution obtained by the stochastic EG-Anderson(1) algorithm outperforms the naive evenly weighted portfolio in terms of both SR and AR.

5.2 Future work

Below are some topics that could be explored in future research work.

1. In this thesis, we incorporate the Anderson(1) acceleration into the design of the proposed algorithms, utilizing its ability to enhance convergence by leveraging information from the two most recent iterations. This approach has proven effective in improving the efficiency and performance of the proposed algorithms. However, there remains significant potential for further advancements. A promising direction for future research is to explore the extension of the EG-Anderson(1) algorithm to EG-Anderson(m) algorithm, which considers

a broader range of historical iteration data. By doing so, we can potentially design even more robust accelerated algorithms that achieve faster convergence rates.

2. Moreover, this thesis focuses on pseudomonotone VIs. For more general non-monotone VIs, such as quasimonotone ones or those satisfying the Minty condition, designing algorithms that ensure global convergence of the entire sequence remains an important direction for future research.

3. In this thesis, we achieve a sublinear convergence rate of $O(N^{-1})$ under the local Lipschitz condition. However, to enhance our results, it is worth exploring additional regularity conditions that could potentially lead to a linear convergence rate. Specifically, the numerical results observed in the EG-Anderson(1) method suggest the possibility of attaining linear convergence under certain circumstances.

Bibliography

[1] H. Abass. Halpern inertial subgradient extragradient algorithm for solving equilibrium problems in Banach spaces. *Appl. Anal.*, 104(2):314–335, 2025.

[2] A. Alacaoglu, A. Böhm, and Y. Malitsky. Beyond the golden ratio for variational inequality algorithms. *J. Mach. Learn. Res.*, 24(172):1–33, 2023.

[3] H. An, X. Jia, and H. F. Walker. Anderson acceleration and application to the three-temperature energy equations. *J. Comput. Phys.*, 347:1–19, 2017.

[4] D. G. Anderson. Iterative procedures for nonlinear integral equations. *J. ACM*, 12(4):547–560, 1965.

[5] H. Attouch, M. O. Czarnecki, and J. Peypouquet. Collective dynamics of higher-order coupled phase oscillators. *SIAM J. Optim.*, 21(4):1251–1274, 2011.

[6] V. Barbu and M. Röckner. Stochastic variational inequalities and applications to the total variation flow perturbed by linear multiplicative noise. *Arch. Ration. Mech. Anal.*, 209(3):797–834, 2013.

[7] W. Bian and X. Chen. Anderson acceleration for nonsmooth fixed point problems. *SIAM J. Numer. Anal.*, 60(5):2565–2591, 2022.

[8] W. Bian, X. Chen, and C. T. Kelley. Anderson acceleration for a class of nonsmooth fixed-point problems. *SIAM J. Sci. Comput.*, 43(5):S1–S20, 2021.

[9] R. I. Boţ, E. R. Csetnek, and P. T. Vuong. The forward-backward-forward method from discrete and continuous perspective for pseudomonotone variational inequalities in Hilbert spaces. *Eur. J. Oper. Res.*, 287:49–60, 2020.

[10] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[11] D. L. Burkholder, B. Davis, and R. F. Gundy. Integral inequalities for convex functions of operators on martingales. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 223–240, 1972.

[12] G. Cai, Q. L. Dong, and Y. Peng. Strong convergence theorems for solving variational inequality problems with pseudomonotone and non-Lipschitz operators. *J. Optim. Theory Appl.*, 188:447–472, 2021.

[13] X. Cai, G. Gu, and B. He. On the $o(1/t)$ convergence rate of the projection and contraction methods for variational inequalities with Lipschitz continuous monotone operators. *Comput. Optim. Appl.*, 57(2):339–363, 2014.

[14] Y. Censor, A. Gibali, and S. Reich. Strong convergence of subgradient extragradient methods for the variational inequality problem in Hilbert space. *Optim. Method Softw.*, 26(4-5):827–845, 2011.

[15] Y. Censor, A. Gibali, and S. Reich. The subgradient extragradient method for solving variational inequalities in Hilbert space. *J. Optim. Theory Appl.*, 148(2):318–335, 2011.

[16] X. Chen and C. T. Kelley. Convergence of the EDIIS algorithm for nonlinear equations. *SIAM J. Sci. Comput.*, 41(1):A365–A379, 2019.

[17] X. Chen and Y. Ye. On homotopy-smoothing methods for box-constrained variational inequalities. *SIAM J. Control Optim.*, 37(2):589–616, 1999.

[18] R. Chugh, N. Gupta, and C. Batra. Modified subgradient extragradient method for solution of split variational inequality problem with application to generalized Nash equilibrium problem. *Optimization*, pages 1–35, 2024.

[19] X. Cui, J. Sun, and L. Zhang. On multistage pseudomonotone stochastic variational inequalities. *J. Optim. Theory Appl.*, 199(1):363–391, 2023.

[20] X. Cui and L. Zhang. A localized progressive hedging algorithm for solving nonmonotone stochastic variational inequalities. *Math. Oper. Res.*, 49(3):1915–1937, 2024.

[21] S. Dafermos and A. Nagurney. A network formulation of market equilibrium problems and variational inequalities. *Oper. Res. Lett.*, 3(5):247–250, 1984.

[22] S. V. Denisov, V. V. Semenov, and L. M. Chabak. Convergence of the modified extragradient method for variational inequalities with non-Lipschitz operators. *Cybern. Syst. Anal.*, 51:757–765, 2015.

[23] Q. L. Dong, Y. J. Cho, L. L. Zhong, and T. M. Rassias. Inertial projection and contraction algorithms for variational inequalities. *J. Glob. Optim.*, 70:687–704, 2018.

[24] R. Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

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

[26] A. Fischer. A Newton-type method for positive-semidefinite linear complementarity problems. *J. Optim. Theory Appl.*, 86:585–608, 1995.

[27] K. Goebel and S. Reich. *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*. Marcel Dekker, New York, 1984.

[28] A. A. Goldstein. Convex programming in Hilbert space. *Bull. Amer. Math. Soc.*, 70:709–710, 1964.

[29] T. N. Hai. Two modified extragradient algorithms for solving variational inequalities. *J. Glob. Optim.*, 78(1):91–106, 2020.

[30] P. T. Harker and J. S. Pang. A damped-Newton method for the linear complementarity problem. *Lectures in Appl. Math.*, 26:265–284, 1990.

[31] P. Hartman and G. Stampacchia. On some nonlinear elliptic differential functional equations. *Acta Math.*, 115(1):153–188, 1966.

[32] C. Heinemann and K. Sturm. Shape optimization for a class of semilinear variational inequalities with applications to damage models. *SIAM J. Math. Anal.*, 48(5):3579–3617, 2016.

[33] N. J. Higham and N. Strabić. Anderson acceleration of the alternating projections method for computing the nearest correlation matrix. *Numer. Algorithms*, 72:1021–1042, 2016.

[34] K. Huang and S. Zhang. Beyond monotone variational inequalities: Solution methods and iteration complexities. *Pac. J. Optim.*, 20(3):403–428, 2024.

[35] K. Huang and S. Zhang. An approximation-based regularized extra-gradient method for monotone variational inequalities. *SIAM J. Optim.*, 35(3):1469–1497, 2025.

[36] A. N. Iusem. An iterative algorithm for the variational inequality problem. *Comput. Appl. Math.*, 13:103–114, 1994.

[37] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.*, 27(2):686–724, 2017.

[38] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM J. Optim.*, 29(1):175–206, 2019.

[39] B. Jadamba and F. Raciti. Variational inequality approach to stochastic Nash equilibrium problems with an application to cournot oligopoly. *J. Optim. Theory Appl.*, 165:1050–1070, 2015.

[40] H. Jiang and H. Xu. Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Trans. Autom. Control*, 53(6):1462–1475, 2008.

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

[42] A. Kannan and U. V. Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Comput. Optim. Appl.*, 74(3):779–820, 2019.

[43] S. Karamardian and S. Schaible. Seven kinds of monotone maps. *J. Optim. Theory Appl.*, 66:37–46, 1990.

[44] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekon. Mat. Metody*, 12:747–756, 1976.

[45] J. Koshal, A. Nedić, and U. V. Shanbhag. Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Trans. Autom. Control*, 58(3):594–609, 2012.

[46] M. Lei and Y. He. An extragradient method for solving variational inequalities without monotonicity. *J. Optim. Theory Appl.*, 188:432–446, 2021.

[47] T. Li, X. Cai, Y. Song, and Y. Ma. Improved variance reduction extragradient method with line search for stochastic variational inequalities. *J. Global. Optim.*, 87(2):423–446, 2023.

[48] G. H. Lin and M. Fukushima. Stochastic equilibrium problems and stochastic mathematical programs with equilibrium constraints: A survey. *Pac. J. Optim.*, 6(3):455–482, 2010.

[49] P. E. Maingé. A hybrid extragradient-viscosity method for monotone operators and fixed point problems. *SIAM J. Control Optim.*, 47(3):1499–1515, 2008.

[50] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM J. Optim.*, 25(1):502–520, 2015.

[51] Y. V. Malitsky and V. Semenov. An extragradient algorithm for monotone variational inequalities. *Cybern. Syst. Anal.*, 50(2):271–277, 2014.

[52] H. Markowitz. Portfolio selection. *J. Finance*, 7(1):77–91, 1952.

[53] W. Ouyang, Y. Liu, and A. Milzarek. Descent properties of an Anderson accelerated gradient method with restarting. *SIAM J. Optim.*, 34(1):336–365, 2024.

[54] W. Ouyang, J. Tao, A. Milzarek, and B. Deng. Nonmonotone globalization for Anderson acceleration via adaptive regularization. *J. Sci. Comput.*, 96(1):5, 2023.

[55] J. Peypouquet and S. Sorin. Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time. *J. Convex Anal.*, 17:1113–1163, 2010.

[56] U. Ravat and U. V. Shanbhag. On the characterization of solution sets of smooth and nonsmooth convex stochastic Nash games. *SIAM J. Optim.*, 21(3):1168–1199, 2011.

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

[58] H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257, 1971.

[59] R. T. Rockafellar and J. Sun. Solving Lagrangian variational inequalities with applications to stochastic programming. *Math. Program.*, 181(2):435–451, 2020.

[60] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Anal.*, 7(4):323–345, 1999.

[61] A. Sripattanet and A. Kangtunyakarn. A new subgradient extragradient method for solving the split modified system of variational inequality problems and fixed point problem. *J. Inequal. Appl.*, 2022(1):51, 2022.

[62] D. V. Thong, V. T. Dung, P. K. Anh, and V. T. Hoang. A single projection algorithm with double inertial extrapolation steps for solving pseudomonotone variational inequalities in Hilbert space. *J. Comput. Appl. Math.*, 426:115099, 2023.

[63] D. V. Thong and A. Gibali. Extragradient methods for solving non-Lipschitzian pseudomonotone variational inequalities. *J. Fixed Point Theory Appl.*, 21:1–19, 2019.

[64] D. V. Thong, Y. Shehu, and O. S. Iyiola. Weak and strong convergence theorems for solving pseudomonotone variational inequalities with non-Lipschitz mappings. *Numer. Algorithms*, 84:795–823, 2020.

[65] D. V. Thong and D. Van Hieu. Modified subgradient extragradient method for variational inequality problems. *Numer. Algorithms*, 79:597–610, 2018.

[66] D. V. Thong, N. T. Vinh, and Y. J. Cho. Accelerated subgradient extragradient methods for variational inequality problems. *J. Sci. Comput.*, 80:1438–1462, 2019.

[67] A. Toth and C. T. Kelley. Convergence analysis for Anderson acceleration. *SIAM J. Numer. Anal.*, 53(2):805–819, 2015.

[68] P. T. Vuong and Y. Shehu. Convergence of an extragradient-type method for variational inequality with applications to optimal control problems. *Numer. Algorithms*, 81(1):269–291, 2019.

[69] W. Wang and B. Ma. Extragradient type projection algorithm for solving quasi-monotone variational inequalities. *Optimization*, 73(8):2639–2655, 2024.

[70] X. Xiao. A unified convergence analysis of stochastic Bregman proximal gradient and extragradient methods. *J. Optim. Theory Appl.*, 188(3):605–627, 2021.

[71] J. Yang and H. Liu. A modified projected gradient method for monotone variational inequalities. *J. Optim. Theory Appl.*, 179:197–211, 2018.

[72] J. Yang and H. Liu. Strong convergence result for solving monotone variational inequalities in Hilbert space. *Numer. Algorithms*, 80:741–752, 2019.

[73] J. Yang, H. Liu, and Z. Liu. Modified subgradient extragradient algorithms for solving monotone variational inequalities. *Optimization*, 67(12):2247–2258, 2018.

[74] Z. P. Yang, J. Zhang, Y. Wang, and G. H. Lin. Variance-based subgradient extragradient method for stochastic variational inequality problems. *J. Sci. Comput.*, 89(1):4, 2021.

[75] Y. Yao, O. S. Iyiola, and Y. Shehu. Subgradient extragradient method with double inertial steps for variational inequalities. *J. Sci. Comput.*, 90:1–29, 2022.

[76] M. Ye. An infeasible projection type algorithm for nonmonotone variational inequalities. *Numer. Algorithms*, 89(4):1723–1742, 2022.

[77] M. Ye and Y. He. A double projection method for solving variational inequalities without monotonicity. *Comput. Optim. Appl.*, 60:141–150, 2015.

[78] F. Yousefian, A. Nedić, and U. V. Shanbhag. Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and Nash games. *IEEE Trans. Autom. Control*, 61(7):1753–1766, 2015.

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

[80] J. Zhang, B. O'Donoghue, and S. Boyd. Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations. *SIAM J. Optim.*, 30(4):3170–3197, 2020.

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