A Modern Treatise on Variational Inequality Problems: Structures, Algorithms, and Iteration Complexities

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kevin Huang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: **Shuzhong Zhang**

May, 2023

Acknowledgements

First, I would like to express my sincerest gratitude to my advisor, Professor Shuzhong Zhang. He has been an amazing and inspiring mentor in both my academic study and my life, and the completion of this dissertation as well as my current achievements would not have been possible without his continual and patient guidance along this path. I could not have imagined a better advisor and mentor during my Ph.D. study, and I am grateful and honored to be his student.

Besides my advisor, I would like to thank the rest of my thesis committee: Professor Zhaosong Lu, Professor William L. Cooper, and Professor Mingyi Hong, for their precious time and insightful comments on my dissertation.

I would also like to thank my colleagues and collaborators: Junyu Zhang and Nuozhou Wang. The discussions among us have been insightful and of great help to my research, and some of the projects in this thesis would not have been completed without these discussions and collaborations. The gratitude extends to the staffs and professors in ISyE. They have provided a warm and supportive environment during my Ph.D. study, and the degree would not have been completed without many of their help.

I am grateful to everyone I met in Minnesota, especially to the friends that I make throughout these years. The time we spent together playing sports, chatting, and having meals is invaluable and makes up an important part of my life here.

I would like to express my greatest gratitude to my family: my parents who have supported me from every aspect and loved me endlessly since I was born; my brother who shares numerous precious memories and enjoyable talks with me; my grandmothers who have always warmly supported me and encouraged me throughout my life.

Finally, special thanks are given to my grandfathers, who have both passed away. They had been so supportive and inspiring in my life, and I could not have imagined myself in today's status without having their loves. I would like to dedicate this dissertation to them.

Abstract

In this thesis, we study the variational inequality (VI) problem with the methodology of designing efficient (or optimal) algorithms and analyzing their (sample/gradient) iteration complexities. In particular, we aim to explore the hidden structures in VI that have not been (fully) studied before and use them as insights to guide the development of new optimal algorithms that align with the modern research trends in both VI and optimization.

We start from the first-order methods, where acceleration has been established in algorithms such as extra-gradient method, optimistic gradient descent ascent method, and dual extrapolation method. These methods are known as optimal in the sense that they match the lower iteration complexity bounds established for first-order methods in (strongly) monotone VI. We observe that these acceleration schemes in VI, together with the acceleration schemes used in optimization such as Nesterov's acceleration, share a common structure: using additional sequence(s), which we refer to it as "extra points", to help improve the convergence of the main sequence. We then propose a general guideline, called the extra-point approach, to construct optimal first-order methods via a more systematic way, which provides flexibility in adopting a variety of extra points/sequences such that the lower bounds take effect. Moving towards high-order methods, research before has relied on using high-order Taylor approximation and an iterative binary-search in solving the subproblems. We show that both of them are not necessary in developing high-order methods, and the key lies in satisfying a high-order Lipschitz bound for any approximation operator used in the subroutine, as well as an appropriate order of regularization to eliminate the needs of binary-search. The proposed unifying framework largely relieves the demand on the complicated analysis derived for different methods and allows us to focus more on the problem structure to design a suitable approximation operator in the algorithm.

We also investigate stochastic algorithms for VI, mainly focusing on the stochastic approximation (SA) approach. We propose stochastic extensions of two new first-order methods, which could be viewed as special instances following the aforementioned extra-point approach, and show that optimal iteration complexities can be established for them, in both situations where the stochastic errors are bounded separately or they are reduced together with the deterministic terms. Application is discussed using the example of black-box saddle point problem where even the function values can only be estimated with noises. Using a smoothing technique, we show that by constructing the stochastic zeroth-order gradients, the previous schemes can be readily applied with the guarantee on sample iteration complexities. Another aspect of stochasticity is discussed following the similar line of research, where we study the VI problems with the finite-sum structure. In addition, such finite-sum structure consists of both general vector mappings and gradient mappings. Developments

in variance reduced algorithms for both finite-sum optimization and finite-sum VI have been found recently, but none has focused on finite-sum VI with optimization structures. We propose two algorithms for both monotone and strongly monotone VI that explicitly make use of such optimization structure and demonstrate that they indeed serve as a bridge between these two problem classes and are able to perform better than general variance reduced VI algorithms when such structure is actually present. We show that the saddle point reformulation of a finite-sum optimization with finite-sum constraints immediately take the aforementioned forms in VI, where applications are commonly seen in Neyman-Pearson classification in machine learning.

Finally, the research in this thesis is extended to non-monotone VI and the solution methods for solving it. Without the monotonicity, another (weaker) global property of the VI problem, the existence of Minty solution, comes to play a central role in the convergence of accelerated projection-type methods. With a slightly worse iteration complexity than the monotone VI, we show that how a general high-order extra-gradient-type method using the concept of the aforementioned approximation can converge. Furthermore, when the existence of Minty solution is no longer assumed, very little can be said in general about the convergence of these projection-type methods. Alternatively, we use these methods as starting points and derive sufficient conditions that characterize various structures of VI where they can converge with guaranteed iteration complexity bounds. This approach allows us to extend our study to potentially broader VI problem classes that have no monotonicity nor Minty solutions.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Literature Review

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set; $F(x) : \mathbb{R}^n \mapsto \mathbb{R}^n$ be a continuous vector mapping. The following problem is known as the "variational inequality (VI) problem":

$$\text{Find } x^* \in \mathcal{X} \text{ such that } F(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in \mathcal{X}. \tag{1.1.1}$$

The study of finite-dimensional VI problems dates back to 1960's where the complementarity problem was developed to solve for various equilibria, such as economic equilibrium, traffic equilibrium, and in general Nash equilibrium. For a comprehensive study of the applications, theories and algorithms of VI, readers are referred to the celebrated monograph by Facchinei and Pang [19].

Throughout this thesis, we are mostly interested in an important class of VI, where the operator F is *monotone*:

$$\langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|^2, \quad \forall x, y \in \mathcal{X} \tag{1.1.2}$$

for some $\mu \geq 0$. If there exists some $\mu > 0$ such that (1.1.2) holds, it is referred to *strongly monotone* and the VI problem has a unique solution. In addition, the Lipschitz continuity is often assumed for the operator F:

$$\|F(x) - F(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathcal{X}, \tag{1.1.3}$$

for some $L \geq \mu$. We denote $\kappa := \frac{L}{\mu}$ as the condition number for VI.

1.1.1 First-order projection methods

The earliest methods developed to solve VI of this type are the projection method due to Sibony [90]:

$$x^{k+1} := \arg\min_{x\in\mathcal{X}} \quad \langle F(x^k), x - x^k\rangle + \frac{\gamma_k}{2}\|x - x^k\|^2, \tag{1.1.4}$$

and the proximal point method due to Martinet [55]:

$$x^{k+1} := \arg\min_{x\in\mathcal{X}} \quad \langle F(x^{k+1}), x - x^k\rangle + \frac{\gamma_k}{2}\|x - x^k\|^2, \tag{1.1.5}$$

for positive $\{\gamma_k\}_{k\geq 0}$, which was later studied and popularized by Rochafellar [82]. These two methods form the basis of most, if not all, methods developed for monotone VI in the research community thus far, in particular for the *projection-type* methods.

Korpelevich [41] first introduced an *extra-step* in the update as follows:

$$\begin{cases} x^{k+0.5} & := & \arg\min_{x\in\mathcal{X}} & \langle F(x^k), x - x^k\rangle + \frac{\gamma_k}{2}\|x - x^k\|^2, \\ \\ x^{k+1} & := & \arg\min_{x\in\mathcal{X}} & \langle F(x^{k+0.5}), x - x^k\rangle + \frac{\gamma_k}{2}\|x - x^k\|^2. \end{cases} \tag{1.1.6}$$

The iteration complexity of the extra-gradient method (1.1.6) is later established by Tseng [96]. In particular, if the operator is strongly monotone ($\mu > 0$), it is $\mathcal{O}\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$ for an ϵ-solution. This is a significant improvement over $\mathcal{O}\left(\kappa^2 \ln\left(\frac{1}{\epsilon}\right)\right)$ of the vanilla projection method (1.1.4), and it is in fact optimal among first-order methods (i.e. using only the information of $F(\cdot)$) applied to such class of problems (with lower bound recently established by Zhang *et al.* [108]). Many algorithms developed for monotone VI thereafter adopt this concept of extra-step update and can be considered as variants of the extra-gradient method, such as modified forward-backward method [97], mirror-prox method [62], dual-extrapolation method [68, 71], hybrid proximal extra-gradient method [60], and the methods generated by the extra-point approach, to be introduced in Chapter 2 of this thesis.

Another type of method, known as *optimistic gradient descent ascent method* (OGDA), was first proposed by Popov [79]:

$$x^{k+1} := P_{\mathcal{X}}\left(x^k - \alpha F(x^k) - \eta(F(x^k) - F(x^{k-1}))\right), \tag{1.1.7}$$

for some positive $\alpha, \eta > 0$, where $P_{\mathcal{X}}$ denotes the projection operator onto $\mathcal{X}$. Unlike the update in extra-gradient method (1.1.6) which uses an extra step, OGDA only requires one update (one projection) per iteration and uses the information from the previous iterate x^{k-1} instead. The optimal convergence of OGDA, in both monotone and strongly monotone VI, is established by Mokhtari *et al.* [58, 59]. The extra-point method proposed by

Huang and Zhang [27] extends the concepts of the extra-gradient method, OGDA, Nesterov's acceleration in optimization [65], and the "heavy-ball" method by Polyak [78] and combines them in a unifying update scheme. If the parameters associated to these different components satisfy a certain constraint set, it is shown that optimal iteration complexity is guaranteed. There is another line of work that studies variants of extra-gradient type methods [103, 47, 39] and proximal point methods [95, 50, 76] with the *anchoring* update, where in each iteration the initial iterate is used as the component of convex combination. The iterates produced are shown to converge among these different methods [104], at a rate same as the optimal convergence rate (to the solution), and the iteration complexities are improved by constant orders compared to vanilla extra-gradient method.

The above methods are known as the *first-order* methods. The lower bound of the iteration complexity for the first-order methods applied to monotone VI is $\Omega\left(\frac{1}{\epsilon}\right)$, as established by Nemirovsky and Yudin [63], while for strongly monotone VI, it is $\Omega\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$, shown by Zhang *et al.* [108] in the context of strongly-convex-strongly-concave saddle-point problems. Methods such as extra-gradient method, mirror-prox method, dual-extrapolation method [68, 71], HPE, OGDA, extra-point method have been proven to achieve these lower bounds, hence optimal.

1.1.2 High-order projection methods

The work of Taji *et al.* [92] is among the first to consider second-order methods for solving VI. A linearized VI subproblem with operator $F(x^k) + \nabla F(x^k)(x - x^k)$ is solved in each iteration and the merit function $f(x) = \max_{x' \in \mathcal{X}} \langle F(x), x - x' \rangle - \frac{\mu}{2}\|x - x'\|^2$ is used to prove the global convergence, with an additional local quadratic convergence. However, no explicit iteration complexity is established for second-order methods until recently. Following the line of research in [92], Huang and Zhang [25] specifically consider unconstrained strongly-convex-strongly-concave saddle point problem and incorporate the idea of cubic regularization (originally proposed by Nesterov in the context of optimization [70]), proving the global iteration complexity $\mathcal{O}\left(\left(\kappa^2 + \frac{\kappa L_2}{\mu}\right) \ln\left(\frac{1}{\epsilon}\right)\right)$, where L_2 is the Lipschitz constant of the Hessian information, in addition to the local quadratic convergence.

Another line of research on second-order methods was started by Monteiro and Svaiter [61]. They propose a Newton Proximal Extragradient (NPE) method, which can be viewed as a special case of the HPE with large step size. In HPE, the first step solves approximately the proximal point update (1.1.5) (denote as $x^{k+0.5}$), while the second step is a regular extra-gradient step. The "large step size" condition, which is key to guarantee a superior

convergence rate, requires:

$$\frac{1}{\gamma_k} \geq \frac{\theta}{\|x^{k+0.5} - x^k\|} \tag{1.1.8}$$

for some constant $\theta > 0$. Note that since $x^{k+0.5}$ depends on γ_k, a certain procedure is required to determine $x^{k+0.5}$ and γ_k such that (1.1.8) also holds. By observing that the set of γ_k satisfying the condition is in fact a closed interval, they develop a bisection method to iteratively reduce the range of γ_k and solve for $x^{k+0.5}$ for each fixed γ_k until the condition is satisfied. They show that for monotone VI, NPE admits $\mathcal{O}\left(1/\epsilon^{\frac{2}{3}}\right)$ iteration complexity for *ergodic mean* of $x^{k+0.5}$ over $0 \leq k \leq N-1$, which is an improvement over the optimal first-order complexity $\mathcal{O}\left(\frac{1}{\epsilon}\right)$. While NPE can also be expressed in the form of second-order mirror-prox method, Bullins and Lai [8] propose a "higher-order mirror-prox method", extending the second-order mirror-prox method to p^{th}-order and establish $\mathcal{O}\left(1/\epsilon^{\frac{2}{p+1}}\right)$ iteration complexity. They replace the linearization $F(x^k) + \nabla F(x^k)(x - x^k)$ with the Taylor approximation of $F(x^k)$, $\sum_{i=0}^{p-1} \frac{1}{i!}\nabla^i F(x^k)[x - x^k]^i$, together with an higher-order constraint on γ_k and $x^{k+0.5}$ similar to (1.1.8). They also demonstrate an explicit procedure to instantiate the proposed method in unconstrained problem with $p = 2$ and a bisection method to search for $x^{k+0.5}$ and γ_k. In [73], Ostroukhov *et al.* further extend the higher-order mirror-prox method to strongly monotone VI by incorporating the *restart procedure*, which yields global iteration complexity $\mathcal{O}\left(\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}} \ln\left(\frac{1}{\epsilon}\right)\right)$. The local quadratic convergence is then guaranteed by incorporating CRN-SPP proposed in [25]. Nesterov in [67] proposes solving constrained convex optimization with cubic regularized Newton method and extends the results to monotone VI with cubic regularized Newton modification of the dual-extrapolation method [68]. The global iteration complexity is shown to be $\mathcal{O}(\frac{1}{\epsilon})$ for monotone VI, with local quadratic convergence for strongly monotone VI.

Recently, there are new developments of higher-order methods for VI. Jiang and Mokhtari [33] propose the Generalized Optimistic Method, which is a general p^{th}-order variant of OGDA. Instead of using $F(x^k)$ to approximate the proximal point update direction $F(x^{k+1})$ with correction $F(x^k) - F(x^{k-1})$ as in OGDA (1.1.7), they propose to use a general approximation $P(x^{k+1}; \mathcal{I}_k)$ with correction $F(x^k) - P(x^k; \mathcal{I}_{k-1})$, where $P(x; \mathcal{I}_k)$ can contain p^{th}-order information and $\mathcal{I}_k$ is the information up to k^{th} iteration. Adil *et al.* [1] propose a pth-order method that improves upon the higher-order mirror-prox method in [8]. The improvement comes from incorporating the gradient of $(p+1)^{\text{th}}$-order regularization in the higher-order VI subproblem, which makes the bisection subroutine unnecessary, and the global complexity is improved by a logarithmic factor. Lin and Jordan [53] propose a pth-order generalization of Nesterov's dual extrapolation method [68], referred to Perseus. Same as [1], Perseus does not require bisection subroutines by solving the VI subproblem with

higher-order regularization. In addition to developing the iteration complexity guarantee in monotone and strongly monotone VI, [53] also extends the analysis to non-monotone VI that satisfies the (strong) Minty condition. Furthermore, they establish the lower bound complexity for general p^{th}-order method applied to monotone VI, given by $\Omega\left(1/\epsilon^{\frac{2}{p+1}}\right)$, which is achieved by Perseus, the Generalized Optimistic Method [33], [1], and ARE to be introduced in Chapter 3 of this thesis. Therefore, they are all optimal p^{th}-order methods for monotone VI.

1.1.3 Stochastic first-order methods

The first-order algorithms for deterministic VI (1.1.1) serve as a basis for the developments of their stochastic counterparts. These algorithms include the aforementioned projection method (1.1.4), the proximal method (1.1.5), the extra-gradient method (1.1.6), the optimistic gradient descent ascent (OGDA) method (1.1.7), the mirror-prox method [62], the extrapolation method [71, 68], and the methods generated from the extra-point approach (see Chapter 2).

In this section, we shall focus on the developments of algorithms for stochastic VI, starting with a paper of Jiang and Xu [32], where the authors propose a stochastic projection method for solving strongly monotone and Lipschitz continuous VI problems and present an almost-sure convergence result. Koshal *et al.* [42] propose iterative Tikhonov regularization method and iterative proximal point method and show almost-sure convergence for the monotone and Lipschitz continuous VI problems. Both methods solve a strongly monotone VI subproblem at each iteration. Yousefian *et al.* [105] further introduce local smoothing technique to the above-mentioned regularized methods to account for non-Lipschitz mappings and show almost-sure convergence. A survey on these methods, as well as applications and the theory behind stochastic VI can be found in Shanbhag [89].

Juditsky *et al.* [35] are among the first to show an iteration complexity bound for stochastic VI algorithms. They extend the mirror-prox method [62] to stochastic settings and prove an optimal iteration complexity bound for monotone VI: $\mathcal{O}(\frac{1}{\epsilon^2})$, or $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ when the variance can be controlled small enough. Yousefian *et al.* [106] further extend the stochastic mirror-prox method with a more general step size choice and show an $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iteration complexity, where they also show an $\mathcal{O}(\frac{1}{\epsilon})$ complexity for the stochastic extra-gradient method for solving strongly monotone VI problems. Yousefian *et al.* [107] use randomized smoothing technique for non-Lipschitz mapping and show an $\mathcal{O}\left(\frac{1}{\epsilon^6}\right)$ iteration complexity. Chen *et al.* [10] consider a specific class of VI model: a mapping that consists of a Lipschitz continuous and monotone operator, a Lipschitz continuous gradient mapping of a convex function, and a subgradient

mapping of a simple convex function. They propose a method that combines Nesterov's acceleration [66] with the stochastic mirror-prox method to exploit this special structure, resulting in an optimal iteration complexity for such class of problem: $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, or $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ when the variance can be controlled small enough, or $\mathcal{O}\left(\sqrt{\frac{1}{\epsilon}}\right)$ when the operator consists only of gradient/subgradient mappings from some convex function. Kannan and Shanbhag [36] analyze a general variant of extra-gradient method (which uses general distance-generating functions) and show that under a slightly weaker assumptions than the strong monotonicity, the optimal $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iteration bound still holds. Kotsalis *et al.* [43] extend the OGDA method to strongly monotone stochastic VI with iteration complexity $\mathcal{O}\left(\max\left\{\frac{\sigma^2}{\mu^2\epsilon}, \kappa \ln \frac{1}{\epsilon}\right\}\right)$.

There have been developments for variance-reduction-based methods in recent years. Jalilzadeh and Shanbhag [31] extend the method [71] for deterministic strongly monotone VI to stochastic VI and show that with variance reduction the optimal iteration complexity $\mathcal{O}(\kappa \ln(1/\epsilon))$ can be achieved, together with a total sample complexity of $\mathcal{O}\left(\frac{1}{\epsilon^\beta}\right)$ for some constant $\beta > 1$. With this method as a subroutine, they also propose a variance-reduced proximal point method with iteration complexity $\mathcal{O}\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$ and sample complexity $\mathcal{O}(\frac{1}{\epsilon^{1+2\alpha\beta}})$ for some constants $\alpha, \beta > 1$. Iusem *et al.* [29] propose a variance-reduced extra-gradient-based method for monotone VI and show $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iteration complexity and $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ sample complexity. They further extend the method [30] by incorporating line-search for unknown Lipschitz constant, while preserving similar bounds. Palaniappan and Bach [74] propose variance-reduced stochastic forward-backward methods based on (accelerated) stochastic gradient descent methods in optimization and show $\mathcal{O}(\kappa \ln(1/\epsilon))$ iteration complexity. In another line of research aiming to model *multistage* stochastic VI (as compared to the *single-stage* VI considered in this paper and most of the afore-mentioned literature), the dynamics between the actions and the arrival of future information play a central role. For the details regarding multistage stochastic VI, the reader is referred to [84, 83] and the references therein. The stochastic oracle may be *man-made*. For instance, the technique of randomized smoothing has been applied in the so-called zeroth-order methods (i.e. derivative-free methods), refer to [72, 87, 17, 88] or the survey [46] in the context of optimization and [99, 101, 54, 85, 56] in the context of minimax saddle-point problems. Note that the optimal sample complexity $\mathcal{O}\left(\frac{d}{\epsilon^2}\right)$ for zeroth-order convex optimization has been established in [17, 88], where d is the problem dimension. On the other hand, similar lower bounds for minimax saddle-point problems have not yet been established, as far as we know.

1.1.4 Finite-sum optimization and VI problems

In machine learning research, a common optimization problem is the so-called finite-sum optimization:

$$\min_{x\in\mathcal{X}} \quad g(x) := \sum_{i=1}^{m} g_i(x), \tag{1.1.9}$$

where the objective is the sum of finitely many (convex) loss functions. When the total number of functions is large, it can be costly for a deterministic gradient method to evaluate the gradients of all the functions in each iteration. A conventional way for solving the finite-sum model (1.1.9) is through stochastic gradient descent (SGD), where in each iteration only one or a mini-batch of functions are randomly chosen and the corresponding gradients are estimated. While SGD may improve the overall gradient complexity over the deterministic methods, the iteration complexity to obtain an ϵ-solution is only $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ even if each of the function $g_i(x)$ is strongly convex and smooth. In order to further improve the gradient and iteration complexity, *variance reduced* algorithms such as SAG [86], SAGA [13], SVRG [34] have been developed to achieve the gradient complexity $\mathcal{O}\left(\left(m+\frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$, assuming each function $g_i(x)$ is strongly convex with modulus $\mu > 0$ and its gradient is Lipschitz continuous with constant $L \geq \mu$. Recently, *accelerated* variance reduced algorithms such as Katyusha [3] and SSNM [109] are proposed to achieve an even better gradient complexity $\mathcal{O}\left(\left(m+\sqrt{\frac{mL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$, which matches the lower bound established in [45], hence optimal.

A specific branch in machine learning which has received much attention in recent years is training Generative Adversarial Network (GAN) [22]. Different from an optimization model (1.1.9), training a GAN can be formulated as a minimax saddle point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \quad f(x,y). \tag{1.1.10}$$

When $f(\cdot,y)$ is convex for fixed $y\in\mathcal{Y}$ and $f(x,\cdot)$ is concave for fixed $x\in\mathcal{X}$ and $\mathcal{X},\mathcal{Y}$ are closed convex sets, (1.1.10) can be reformulated into the VI model (1.1.1). Since GAN is known to be very difficult to train and the conventional (stochastic) gradient methods applied for deep learning do not perform well in practice, there has been a surge of interest in developing efficient gradient methods in the context of either saddle point problem or VI [11, 57, 49, 21, 26]. It is also natural to consider the finite-sum VI where $F(x) = \sum_{i=1}^{m} F_i(x)$ and develop variance reduced algorithms applying techniques from finite-sum optimization. The authors in [2] incorporated such variance reduced techniques into various VI algorithms and established the gradient complexity $\mathcal{O}\left(m+\frac{\sqrt{m}L}{\epsilon}\right)$ for monotone VI and $\mathcal{O}\left(\left(m+\frac{\sqrt{m}L}{\mu}\right)\log\frac{1}{\epsilon}\right)$ for strongly monotone VI, where each operator $F_i(x)$ is (strongly) monotone with modulus $\mu\,(>)\geq 0$ and Lipschitz continuous with $L\geq\mu$. On the other hand,

a lower gradient complexity bound has also been established in [100] with $\Omega\left(m + \frac{L}{\epsilon}\right)$ for convex-concave saddle point problem and $\Omega\left(\left(m + \frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$ for strongly-convex-strongly-concave saddle point problem. Unlike the accelerated variance reduced algorithms in optimization [3, 109] which have been proven to be optimal, there is still a gap between the upper and lower gradient complexity bounds for finite-sum VI. It remains an open problem to determine where the optimal gradient complexity bound actually lands.

1.2 Overview and Organization

In Chapter 2, we propose an extra-point approach, which is used as a general iterative procedure to guide the developments of optimal first-order algorithms. To illustrate the idea, we present three different schemes whose forms follow the general extra-point approach, for solving strongly monotone VI, strongly monotone VI in combination with monotone gradient mapping, and strongly convex optimization. Iteration complexity analysis shows that these schemes are optimal within respective problem classes.

In Chapter 3, we propose a unifying framework for general p^{th}-order methods in VI, which we call Approximation-based Regularized Extra-gradient method (ARE). We first conduct iteration complexity analysis for global convergence for both monotone and strongly monotone VI, and we further show that local superlinear convergence of order p can be established for strongly monotone VI with a modified ARE algorithm that retains the previous global convergence guarantee. Discussions are further extended to solving high-order VI subproblems in ARE iterations with special instances of approximation operators, as well as examples of structured ARE schemes when the operator is of the composite form, $F = H + G$, or more generally $F = H \cdot G$.

In Chapter 4, we study stochastic methods for VI. We consider the stochastic approximation approach and propose stochastic extensions of two new first-order algorithms, the stochastic extra-point scheme and the stochastic extra-momentum scheme. We show that they both achieve the optimal iteration complexity $\mathcal{O}\left(\kappa \ln\frac{1}{\epsilon}\right)$ with additional stochastic error terms, or $\mathcal{O}\left(\max\left\{\kappa\ln\frac{1}{\epsilon}, \frac{\sigma^2}{\epsilon\mu^2}\right\}\right)$ if we reduce the stochastic errors simultaneously throughout the iterations. We then introduce a stochastic black-box saddle-point problem as a specific application to stochastic VI and present the sample complexity results the proposed methods.

In Chapter 5, we develop stochastic variance reduced algorithms for VI problems with finite-sum general vector mappings and finite-sum gradient mappings. There are two algorithms, SAVREP and SAVREP-m, proposed for strongly monotone VI and monotone VI respectively, sharing the common high-level structures with different treatments in both specific

implementation and gradient complexity analysis. We demonstrate the applications of such finite-sum VI structure for solving finite-sum convex optimization with finite-sum inequality constraints, with an extension to the black-box setting.

In Chapter 6, we research on non-monotone VI. The discussion starts with non-monotone VI with Minty solution, where we first provide formal definitions of the solution concepts and merit functions that are relevant to the discussion in this chapter. Then the convergence of ARE is established, followed by discussions on implications of Minty solutions in optimization and Nash games. Finally, we explore algorithm-based sufficient conditions and establish convergence of gradient-projection method and extra-gradient method under these conditions.

Chapter 2

Optimal First-Order Schemes for VI Problems and Convex Optimization: an Extra-Point Approach

In this chapter, we propose an extra-point approach for constructing flexibly trainable optimal first-order schemes for a given VI and/or optimization problem class. As opposed to the main iterates whose convergence is of concern, the *extra points* form additional sequences generated to improve the convergence of the main iterates. To illustrate the idea, we present three such solution schemes of first-order algorithms for strongly monotone VI, strongly monotone VI in combination with monotone gradient mapping, and strongly convex optimization, respectively. In the proposed schemes, the main iterates and extra points are generated within the scope where the lower bounds on the iteration complexity ([66, 108]) take effect. While these schemes follow the general extra-point approach, the specific updating procedures leverage on the search directions suggested by the well-established accelerated first-order methods, such as OGDA, the heavy-ball method, the extra-gradient method, and Nesterov's acceleration. This demonstrates concrete ways to develop first-order methods which not only match the lower complexity bounds (hence *optimal*) but also introduce flexibilities allowing performance improvement and enhancement tailored for each problem class at hand.

2.1 Introduction

2.1.1 Problem of interest

In this chapter, we propose to use an *extra-point* approach to develop general optimal first-order algorithmic schemes. To illustrate the idea, we specifically discuss three different problem classes, which we shall briefly introduce here to set the stage. The more in-depth discussions for each problem class are presented in Sections 2.3, 2.4, and 2.5 respectively, where the detailed update procedures of the schemes are also formally introduced.

The first problem class of interest is the variational inequality (VI) problem. Since in this chapter, the discussions are conducted for different problems classes rather than limited to just VI, we present the problem formulation here once again for the sake of clear comparisons and referencing. Given a constraint set $\mathcal{Z} \subseteq \mathbb{R}^n$ and a mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$, find $z^* \in \mathcal{Z}$ such that

$$F(z^*)^\top (z - z^*) \geq 0, \quad \forall z \in \mathcal{Z}. \tag{2.1.1}$$

In this chapter, we consider the VI model (2.1.1) where $\mathcal{Z}$ is a closed convex set. Moreover, throughout this chapter we assume the strong monotonicity (1.1.2) and Lipschitz continuity (1.1.3) of F. With the different notation of variables used in this chapter, they are summarized as follows once again:

$$\big(F(z) - F(z')\big)^\top (z - z') \geq \mu \|z - z'\|^2, \quad \forall z, z' \in \mathcal{Z}, \tag{2.1.2}$$

for some $\mu > 0$, and

$$\|F(z) - F(z')\| \leq L\|z - z'\|, \quad \forall z, z' \in \mathcal{Z}, \tag{2.1.3}$$

for some $L \geq \mu > 0$.

The second problem class of interest is an extended class of VI (2.1.1) where the mapping F can be expressed as the sum of a general vector mapping $H(z)$ and a gradient mapping $\nabla g(z)$:

$$F(z) := H(z) + \nabla g(z). \tag{2.1.4}$$

This problem is studied in [10] in the stochastic setting when $H(z)$ is monotone and $g(z)$ is convex. The presence of the gradient mapping $\nabla g(z)$ enables a more accurate estimation on the overall continuity of $F(z)$, and the authors of [10] show that the proposed accelerated mirror-prox method can solve this problem faster than optimal approaches developed for general VI (2.1.1) and in fact meets the lower bound (see Section 2.4). In this chapter, we

focus on the deterministic problem and extend the study to the case when $H(z)$ is strongly monotone (2.1.2).

The last problem class of interest is the optimization model:

$$\min_{x \in \mathcal{X}} f(x), \tag{2.1.5}$$

where the first-order optimality condition given f is smooth, convex, and $\mathcal{X}$ is a convex set:

$$\nabla f(x^*)^\top (x - x^*) \geq 0, \quad \forall x \in \mathcal{X}$$

has the same formulation as (2.1.1) when F is a gradient mapping. Therefore, it is also known as one of the most important applications of VI (and also a special case of (2.1.4) when $H(z) = 0$). However, with the special structure in optimization, the algorithms are in general faster than those designed for the previous two classes of problems. In this chapter, we assume that $\mathcal{X} := \mathbb{R}^n$, i.e. the problem is unconstrained, and $f(x)$ is strongly convex with modulus μ and the gradient Lipschitz continuous with constant L.

In the rest of this chapter, we often use z to denote the variables in the VI-related context, while using x to denote the variables in the optimization-related context, in order to differentiate between the two different problem classes. Also denote $\kappa := \frac{L}{\mu} \geq 1$ and $\sigma := \frac{\mu}{L} = \frac{1}{\kappa} \leq 1$. Parameter κ is also known as the condition number of problems (2.1.1), (2.1.4), and (2.1.5).

2.1.2 The lower iteration complexity bounds

Given a precision $\epsilon > 0$, while the *upper-bound* analysis for an algorithm is concerned with the worst-case iteration complexity required to reach an ϵ-solution (formally defined for different problem classes in later discussion), the research in *lower-bound* analysis of a problem class is concerned with the best iteration complexity bound that can be achieved within a class of algorithms. The classic reference on the information-theoretic results for lower iteration complexity bounds is Nemirovski and Yudin [63]. In Nemirovski [64], the lower complexity bound for convex quadratic optimization is established. Discussions on the iteration complexity lower bounds for general convex optimization $\Omega(1/\sqrt{\epsilon})$ and for strongly convex optimization $\Omega(\sqrt{\kappa}\ln(1/\epsilon))$ can be found in Nesterov's monograph [66]. The most recent development regarding the iteration complexity lower bound for optimization is in [15], where the authors establish an exact lower bound for strongly convex optimization, in the sense that it is achieved up to a constant by ITEM [93].

Recently, in Zhang *et al.* [108] the authors consider strongly-convex-strongly-concave saddle point problems, and show that there exists a class of problems in such a way that no first-

order algorithm can find an ϵ-solution in less than

$$\Omega\left(\sqrt{\frac{L_x}{\mu_x}+\frac{L_{xy}^2}{\mu_x\mu_y}+\frac{L_y}{\mu_y}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

iterations. Since the strongly-convex-strongly-concave saddle point problem is a special class of strongly monotone VI with $L_x = L_y = L_{xy} = L$ and $\mu_x = \mu_y = \mu$, the above lower bound can be regarded as a lower bound result for general strongly monotone VI problems, given by $\Omega\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$. This is in sharp contrast to the case of strongly convex optimization, where the lower bound for the iteration complexity is $\Omega\left(\sqrt{\kappa}\ln\left(\frac{1}{\epsilon}\right)\right)$. For an extended class of VI given in the form (2.1.4), the lower complexity bounds can be derived by combining the results from VI and optimization. We shall postpone a detailed discussion until Section 2.4; see also the discussions in [10].

2.1.3 An extra-point approach to optimal first-order algorithms

In the development of the lower bounds [66, 108], the class of algorithms is restricted to the so-called first-order methods; that is, the subspace formed by all past iterates and the vectors formed by their first-order information (gradient for optimization and general vector mapping for VI). The algebraic manipulations among these information are therefore affine linear, including proximal mappings. An algorithm within this class is said to be *optimal* if its iteration complexity matches the order of magnitude of the lower bound. However, it is in general difficult to determine certain forms of first-order methods, or more importantly, optimal first-order methods, given the aforementioned broadly defined class of first-order algorithms. In this chapter, we aim to provide general guidelines for developing optimal first-order methods within this class. We introduce the concept of extra points, which plays a key role in our framework. The proposed general iterative procedure based on this concept is described as follows, which we shall refer to as the *extra-point approach*:

Step 1 Obtain first-order information of a subset of the current *main iterate sequence* and the *extra-point sequences*.

Step 2 Perform linear manipulations on a subset of the current sequences and the first-order information obtained in **Step 1** to generate *search directions*.

Step 3 Update the main iterates and the extra points with the search directions obtained in **Step 2**.

A few remarks are in order here to clarify the above extra-point approach. First, we classify the sequences generated throughout the algorithm into two groups: the main iterate

sequence and the extra-point sequences. The main iterate sequence (or simply *the main iterates* henceforth) consists of iterates whose convergence to the solution set is of main concern, and the iteration complexity of an algorithm is determined by the number of iterations required when the main iterates converge to an ϵ-solution. On the other hand, the extra-point sequences (or simply *the extra points* henceforth) are the iterates generated to take part in the update of the main-iterate sequence, which often play an important role in achieving optimal iteration complexity. While the extra-point sequences may also converge, it is in general not of concern in the analysis. Secondly, the *search directions* play a central role in updating the sequences. They are often obtained by affine linear combinations of the main iterates, the extra points, and the first-order information thereof. One of the most commonly used search directions is the *gradient descent* direction, which is the negation of the gradient of the current main iterate. On the other hand, using search directions involving extra points are often shown to be critical in the design of optimal algorithms, which we shall elaborate with more details in the later sections. Note that the update of certain extra points may only involve linear combinations of the extra points and the main iterates (such as certain convex combinations or extrapolations), while others may involve using search directions. In contrast, the updates of the main iterates usually rely on using (multiple) search directions. The above described characteristics of the extra-point approach provide a general guideline for developing (optimal) first-order algorithms within the broadly defined class where the lower bound results are effective [66, 108]. In this chapter, we show that the extra-point approach can in fact lead us to construct optimal algorithms for VI and convex optimization, while maintaining a great deal of (yet manageable) flexibilities suitable for potential parameter-tuning and structure-learning, which is critically important for the success of any first-order algorithm. We present three such case studies in Sections 2.3, 2.4 and 2.5.

2.1.4 Accelerated first-order methods

In this section we introduce the main constituents of the exemplifying schemes which will be discussed in later sections. In particular, they use extra points and search directions suggested by some existing *accelerated* first-order algorithms. In 1983 [65], Nesterov proposed an algorithm for convex optimization with the iteration complexity $\mathcal{O}(1/\sqrt{\epsilon})$ to reach an ϵ-solution, known as "Nesterov's accelerated method". Since Nesterov's accelerated method matches the order of the lower bound for smooth convex optimization, it is also known as an *optimal* algorithm. The term "acceleration" is then used in contrast of the vanilla gradient descent method, which only yields a suboptimal iteration complexity $\mathcal{O}(1/\epsilon)$. In the context of strongly convex optimization, Nesterov's method can be modified to yield the optimal

$\mathcal{O}(\sqrt{\kappa}\ln(1/\epsilon))$ iteration complexity [66]. There has been an intensive recent research effort on the subject; see the recent monograph [12] for a comprehensive survey. We devote Section 2.5 to a technical discussion on the subject in the context of optimization. At this point, we shall mention another accelerated method that predates Nesterov's acceleration, which leverages on the so-called momentum of the dynamics in the gradient fields. The method was introduced by Polyak in 1964 [78] and is commonly known by the name of the "heavy-ball" method, which can be shown to yield an iteration complexity $\mathcal{O}(\sqrt{\kappa}\ln(1/\epsilon))$ for minimizing a strongly convex quadratic objective function.

On the side of algorithmic design for VI, classical methods include the projection algorithm by Sibony [90], the proximal point method as proposed by Martinet [55] and popularized by Rockafeller [82], and the matrix splitting method of Tseng [96]. In this section we shall focus on the methods involving projection of iterates onto the feasible set. In the context of strongly monotone VI, vanilla projection method yields a suboptimal complexity bound of $\mathcal{O}(\kappa^2\ln(1/\epsilon))$, and the "accelerated" methods include the following developments. The so-called extra-gradient method was proposed by Korpelevich in 1976 [41] for the saddle point problems, which was shown to be linearly convergent for strongly monotone VI by Tseng in [96]. The optimistic gradient descent ascent (OGDA) method was proposed by Popov in 1980 [79]. That method has a close relation with the so-called momentum-based methods, and we shall come back to this point later. The complexity of OGDA was studied by Mokhtari *et al.* in [59] for convex-concave saddle point problem and by Palaniappan and Bach [74] and Mokhtari *et al.* [58] for strongly monotone VI. Nesterov proposed a dual-extrapolation method for monotone VI [68] and for strongly monotone VI [71]. The above mentioned extra-gradient method, OGDA, and dual-extrapolation method achieve an iteration complexity of $\mathcal{O}(\kappa\ln(1/\epsilon))$. Noting that the minimax saddle point model can be formulated as VI, in view of the lower bound established in Zhang *et al.* [108] those algorithms all have reached the lower bound. Therefore, they are all optimal algorithms for solving strongly monotone VI.

2.2 Acceleration with Extra Points

2.2.1 Vanilla gradient projection and the extra-gradient method

In this section we shall conduct an analysis revealing the mechanism leading to various phenomena of acceleration; i.e., incorporating extra points when updating the main iterates. While the exact mechanism in VI and optimization can be quite different, they arguably share similar underlying ideas, which motivate the proposed extra-point approach and the

specialized forms adopted in the exemplifying schemes. We shall focus our discussion on strongly monotone VI in this section, while leaving the discussion of strongly convex optimization to Section 2.5.

Let us first introduce the projection operator $P_{\mathcal{Z}}(\cdot)$

$$P_{\mathcal{Z}}(z) = \arg\min_{z' \in \mathcal{Z}} \|z - z'\|^2 \tag{2.2.6}$$

and its first-order optimality condition:

$$\langle P_{\mathcal{Z}}(z) - z, z' - P_{\mathcal{Z}}(z)\rangle \geq 0, \quad \forall z' \in \mathcal{Z},$$

which we will use throughout the analysis in this paper.

The vanilla projection method updates the main iterates as follows:

$$z^{k+1} = P_{\mathcal{Z}}\left(z^k - \alpha F(z^k)\right), \quad k = 0, 1, 2, ...$$

The first-order optimality condition is given by:

$$\langle z^{k+1} - (z^k - \alpha F(z^k)), z - z^{k+1}\rangle \geq 0, \quad \forall z \in \mathcal{Z}.$$

Take $z = z^*$ and rearrange the terms:

$$\begin{aligned}
& \frac{1}{2}\left(\|z^{k+1} - z^k\|^2 + \|z^{k+1} - z^*\|^2 - \|z^k - z^*\|^2\right) \leq \alpha\langle F(z^k), z^* - z^{k+1}\rangle \\
= \;& \alpha\langle F(z^k) - F(z^*), z^* - z^{k+1}\rangle + \langle F(z^*), z^* - z^{k+1}\rangle \leq \alpha\langle F(z^k) - F(z^*), z^* - z^{k+1}\rangle \\
= \;& \alpha\langle F(z^k) - F(z^*), z^* - z^k\rangle + \alpha\langle F(z^k) - F(z^*), z^k - z^{k+1}\rangle \\
\leq \;& -\alpha\mu\|z^k - z^*\|^2 + \alpha L\|z^k - z^*\|\|z^{k+1} - z^k\| \\
\leq \;& \frac{1}{2}\left(-2\alpha\mu + \alpha^2 L^2\right)\|z^k - z^*\|^2 + \frac{1}{2}\|z^{k+1} - z^k\|^2.
\end{aligned}$$

We then get

$$\|z^{k+1} - z^*\|^2 \leq (1 - 2\alpha\mu + \alpha^2 L^2)\|z^k - z^*\|^2 \overset{\alpha := \frac{\mu}{L^2}}{=} (1 - \sigma^2)\|z^k - z^*\|^2.$$

Note that the term $\alpha\langle F(z^k) - F(z^*), z^* - z^k\rangle$ would guarantee a linear rate $1 - \alpha\mu$ if there were no $\alpha^2\|F(z^k) - F(z^*)\|^2$ term. The presence of the latter term causes a smaller step size of $\alpha = \frac{\mu}{L^2}$, leading to a reduction rate of $1 - \sigma^2$.

The idea behind the extra-gradient method is to introduce an extra-point sequence $\{z^{k+0.5}\}$ to evaluate the gradient so as to avoid dealing with the latter term as such. The extra-gradient method proceeds as follows:

$$\left\{\begin{array}{rcl} z^{k+0.5} & = & P_{\mathcal{Z}}\left(z^k - \alpha F(z^k)\right), \\ z^{k+1} & = & P_{\mathcal{Z}}\left(z^k - \alpha F(z^{k+0.5})\right), \quad k = 0, 1, 2, ... \end{array}\right. \tag{2.2.7}$$

The optimality condition of $z^{k+0.5}$ is given by:

$$\langle z^{k+0.5} - (z^k - \alpha F(z^k)), z - z^{k+0.5} \rangle \geq 0, \quad \forall z \in \mathcal{Z},$$

which can be equivalently written as

$$\langle \alpha F(z^k), z - z^{k+0.5} \rangle \geq \frac{1}{2} \left(\|z^{k+0.5} - z^k\|^2 + \|z^{k+0.5} - z\|^2 - \|z - z^k\|^2 \right), \quad \forall z \in \mathcal{Z}. \tag{2.2.8}$$

On the other hand, the optimality condition of z^{k+1} is given by:

$$\langle z^{k+1} - (z^k - \alpha F(z^{k+0.5})), z - z^{k+1} \rangle \geq 0, \quad \forall z \in \mathcal{Z},$$

which results in

$$\langle \alpha F(z^{k+0.5}), z - z^{k+1} \rangle \geq \frac{1}{2} \left(\|z^{k+1} - z^k\|^2 + \|z^{k+1} - z\|^2 - \|z - z^k\|^2 \right), \quad \forall z \in \mathcal{Z}. \tag{2.2.9}$$

Continuing from (2.2.9), it follows that

$$\begin{aligned}
& \frac{1}{2} \left(\|z^{k+1} - z^k\|^2 + \|z^{k+1} - z\|^2 - \|z - z^k\|^2 \right) \leq \langle \alpha F(z^{k+0.5}), z - z^{k+1} \rangle \\
= \; & \langle \alpha F(z^{k+0.5}), z - z^{k+0.5} \rangle + \langle \alpha F(z^{k+0.5}), z^{k+0.5} - z^{k+1} \rangle \\
= \; & \langle \alpha F(z^{k+0.5}), z - z^{k+0.5} \rangle + \alpha \langle F(z^{k+0.5}) - F(z^k), z^{k+0.5} - z^{k+1} \rangle \\
& + \alpha \langle F(z^k), z^{k+0.5} - z^{k+1} \rangle \\
\leq \; & \langle \alpha F(z^{k+0.5}), z - z^{k+0.5} \rangle + \alpha \|F(z^{k+0.5}) - F(z^k)\| \|z^{k+0.5} - z^{k+1}\| \\
& + \alpha \langle F(z^k), z^{k+0.5} - z^{k+1} \rangle \\
\leq \; & \langle \alpha F(z^{k+0.5}), z - z^{k+0.5} \rangle + \frac{\alpha^2}{2} \|F(z^{k+0.5}) - F(z^k)\|^2 + \frac{1}{2} \|z^{k+0.5} - z^{k+1}\|^2 \\
& + \alpha \langle F(z^k), z^{k+0.5} - z^{k+1} \rangle \\
\leq \; & \langle \alpha F(z^{k+0.5}), z - z^{k+0.5} \rangle + \frac{\alpha^2 L^2}{2} \|z^{k+0.5} - z^k\|^2 + \frac{1}{2} \|z^{k+0.5} - z^{k+1}\|^2 \\
& + \alpha \langle F(z^k), z^{k+0.5} - z^{k+1} \rangle.
\end{aligned}$$

We shall use (2.2.8) with $z = z^{k+1}$ to bound the last term of the above inequality. By cancelling out terms, we get:

$$\|z^{k+1} - z\|^2 - \|z^k - z\|^2 \leq 2\alpha \langle F(z^{k+0.5}), z - z^{k+0.5} \rangle + (\alpha^2 L^2 - 1) \|z^{k+0.5} - z^k\|^2,$$

which holds for all $z \in \mathcal{Z}$. Substituting in $z = z^*$ and noting that

$$\begin{aligned}
& \langle F(z^{k+0.5}), z^* - z^{k+0.5} \rangle = \langle F(z^{k+0.5}) - F(z^*), z^* - z^{k+0.5} \rangle + \langle F(z^*), z^* - z^{k+0.5} \rangle \\
\leq \; & \langle F(z^{k+0.5}) - F(z^*), z^* - z^{k+0.5} \rangle \leq -\mu \|z^{k+0.5} - z^*\|^2 \\
\leq \; & -\frac{\mu}{2} \|z^k - z^*\|^2 + \mu \|z^{k+0.5} - z^k\|^2,
\end{aligned} \tag{2.2.10}$$

we obtain

$$\begin{aligned}\|z^{k+1}-z^*\|^2 &\leq (1-\alpha\mu)\|z^k-z^*\|^2+(\alpha^2L^2+2\alpha\mu-1)\|z^k-z^{k+0.5}\|^2\\ &\overset{\alpha:=\frac{1}{4L}}{\leq} \left(1-\frac{\sigma}{4}\right)\|z^k-z^*\|^2.\end{aligned}$$

We can see that by introducing an extra-gradient step of $F(z^{k+0.5})$, the upper bound estimate from the Lipschitz constant L is now absorbed to the term $\|z^k-z^{k+0.5}\|^2$ instead of $\|z^k-z^*\|^2$ in vanilla projection method, which allows the extra-gradient method to take a more aggressive step size $\alpha=\frac{1}{4L}$, leading to an improved reduction rate $1-\Omega(\sigma)$.

2.2.2 Other accelerated first-order methods

We introduce other accelerated first-order methods whose search directions and/or extra points will be adopted by the exemplifying schemes to follow. The detailed analysis is omitted here. The first method (OGDA) is used to solve general monotone VI, while the other two methods (Nesterov's acceleration and heavy-ball method) are for convex optimization.

The optimistic gradient descent ascent (OGDA) method

Unlike extra-gradient method, which updates iterate k with the mapping at $z^{k+0.5}$, the OGDA method updates with an extrapolated mapping direction:

$$z^{k+1}=P_{\mathcal{Z}}\left(z^k-\alpha F(z^k)-\tau\left(F(z^k)-F(z^{k-1})\right)\right). \tag{2.2.11}$$

For parameter choices $\alpha=\frac{1}{2L}$ and $\tau=\frac{\alpha}{1+\sigma}$, OGDA yields a linear convergence:

$$\|z^k-z^*\|^2\leq 2(1+\sigma)^{-k}\|z^0-z^*\|^2.$$

For interested readers, we refer the proof to [74, 58]. Note that the term $F(z^k)-F(z^{k-1})$ is known as the *optimism*, initially proposed by Popov [79].

The heavy-ball method

The heavy-ball method proposed in [78] was designed to solve strongly convex optimization $\min_x f(x)$ with the update rule:

$$x^{k+1}=x^k-\frac{4}{(\sqrt{L}+\sqrt{\mu})^2}\nabla f(x^k)+\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)(x^k-x^{k-1}). \tag{2.2.12}$$

The heavy-ball method is known to have an improved rate of convergence as compared to the regular gradient method (see [78]), if f is a strongly convex quadratic function. In particular, we have

$$\left\| \begin{pmatrix} x^{k+1} - x^* \\ x^k - x^* \end{pmatrix} \right\| \leq C \cdot \left(\frac{1-\sqrt{\sigma}}{1+\sqrt{\sigma}} + \delta_k \right)^k \cdot \left\| \begin{pmatrix} x^k - x^* \\ x^{k-1} - x^* \end{pmatrix} \right\|, \tag{2.2.13}$$

where C is a constant independent of σ and $\lim_{k\to\infty} \delta_k = 0$.

Nesterov's method

Nesterov's accelerated gradient method for strongly convex optimization $\min_x f(x)$ can be stated as follows [66]:

$$x^{k+1} = x^k + \beta(x^k - x^{k-1}) - \alpha \nabla f(x^k + \beta(x^k - x^{k-1})),$$

where $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. Note that the above updating formula is for strongly convex minimization. (If the function is merely convex, we then let β be depending on the iteration count k: $\beta_k = \frac{k}{k+3}$.) Nesterov's method for strongly convex minimization has a linear convergence rate as follows:

$$f(x^k) - f(x^*) \leq 2(1-\sqrt{\sigma})^k \left(f(x^0) - f(x^*) \right).$$

2.2.3 The extra-point formulations

Observe that in Section 2.2.2, the updating rules for the three methods follow a similar logic. In general, the updating formula can be expressed as generating an extra-point sequence $\{x^{k+0.5}\}$ first, which is used in the update of the main iterate sequence $\{x^k\}$. For OGDA, if we assume $\mathcal{Z} = \mathbb{R}^n$ and $F(\cdot)$ is *linear*, then the updates can be written as:

$$\begin{cases} z^{k+0.5} &= z^k + \beta(z^k - z^{k-1}), \\ z^{k+1} &= z^k - \alpha F(z^{k+0.5}). \end{cases}$$

The heavy-ball method is in the form:

$$\begin{cases} x^{k+0.5} &= x^k + \beta(x^k - x^{k-1}), \\ x^{k+1} &= x^{k+0.5} - \alpha \nabla f(x^k). \end{cases}$$

Nesterov's accelerated method is in the form:

$$\begin{cases} x^{k+0.5} &= x^k + \beta(x^k - x^{k-1}), \\ x^{k+1} &= x^{k+0.5} - \alpha \nabla f(x^{k+0.5}). \end{cases}$$

Obviously, the exact way how these extra points are combined and how the main iterates are updated (i.e. the parameter choices) could largely vary from one to another depending on the specific problem classes under consideration. However, the common underlying idea is clear: that is, to generate extra-point sequences other than the main iterate sequence that take part in the update of the latter in certain ways. While the convergence of the extra-point sequences may not be the main concern, the accelerated convergence rate of the main iterate sequences is eventually achieved with the presence of these extra points. This idea is therefore extracted as the guiding principle in the design of our newly proposed extra-point schemes, which we shall put forward in the next three sections.

2.3 An Extra-Point Approach to General Strongly Monotone VI Problems

Our proposed extra-point scheme for solving the strongly monotone VI model (2.1.1) is based on the following updating formula:

$$\left\{\begin{array}{rcl} z^{k+0.5} & = & P_{\mathcal{Z}}\left(z^k + \beta(z^k - z^{k-1}) - \eta F(z^k)\right), \\ z^{k+1} & = & P_{\mathcal{Z}}\left(z^k - \alpha F(z^{k+0.5}) + \gamma(z^k - z^{k-1}) - \tau\left(F(z^k) - F(z^{k-1})\right)\right) \end{array}\right. \tag{2.3.14}$$

for $k = 0, 1, 2, ...$ and $z^{-1} := z^0$.

Similar to the extra-gradient method (2.2.7), the above scheme introduces an extra-point sequence $\{z^{k+0.5}\}$ to help with the update of the main iterate sequence $\{z^k\}$. Furthermore, it adopts additional search directions from other accelerated first-order methods in the update of $z^{k+0.5}$ and z^{k+1}. In order for these search directions to collaborate such that the optimal iteration complexity is guaranteed, the scheme (2.3.14) requires five nonnegative parameters to operate with: $\alpha, \beta, \gamma, \tau, \eta$, where α and η can be related to gradient/extra-gradient search directions; β and γ can be related to Nesterov's acceleration/heavy-ball search directions; τ can be related to optimism.

While the proposed (2.3.14) is merely one of the many possible specific forms that an extra-point approach can take for solving strongly monotone VI, it is already flexible enough to be considered as a more general scheme than some existing accelerated methods. By taking specific parameter configurations, (2.3.14) will reduce to simpler forms, where the dynamics are summarized in Table 2.1. Note that we have omitted the projection operator (or $\mathcal{Z} := \mathbb{R}^n$) for clearer comparison. While the particular methods in Table 2.1 are all optimal in solving certain problem classes, we shall show in the rest of this section that the proposed scheme (2.3.14) is also optimal if the parameters are within a given range of configurations.

Existing Method	α	β	η	γ	τ	*The Dynamics*
vanilla projection	+	0	0	0	0	$z^{k+1} = z^k - \alpha F(z^k)$
"heavy-ball"	+	0	0	+	0	$z^{k+1} = z^k - \alpha F(z^k) + \gamma(z^k - z^{k-1})$
extra gradient	+	0	+	0	0	$z^{k+1} = z^k - \alpha F(z^k - \eta F(z^k))$
Nesterov's method	+	+	0	+	0	$z^{k+1} = z^k - \alpha F(z^k + \beta(z^k - z^{k-1})) + \gamma(z^k - z^{k-1})$
OGDA	+	0	0	0	+	$z^{k+1} = z^k - \alpha F(z^k) - \tau(F(z^k) - F(z^{k-1}))$

Table 2.1: Parameters correspondence in different first-order methods

To analyze iteration complexity of scheme 2.3.14, let us first establish the following relation:

Lemma 2.3.1. *For the sequences* $\{z^k\}$ *and* $\{z^{k+0.5}\}$, $k = 0, 1, 2, \ldots$ *generated from the scheme* (2.3.14)*, the following inequality holds*

$$\begin{aligned} & (1-\tau L)\|z^{k+1} - z^*\|^2 \\ \leq\ & (1 - \alpha\mu + 4\gamma + 2|\gamma-\beta| + 2\tau L)\|z^k - z^*\|^2 + (2\gamma + 2|\gamma-\beta| + 2\tau L)\|z^{k-1} - z^*\|^2 \\ & +(|\gamma-\beta| - 1/2)\|z^{k+1} - z^{k+0.5}\|^2 + (2\alpha^2L^2 + 2\alpha\mu + 2\gamma - 1)\|z^{k+0.5} - z^k\|^2 \\ & +2(\eta-\alpha)F(z^k)^\top(z^{k+1} - z^{k+0.5}). \end{aligned} \tag{2.3.15}$$

Proof. We shall show that the following inequality can be established:

$$\begin{aligned} & 2\alpha\langle F(z^{k+0.5}), z^{k+0.5} - z\rangle + (1-\tau L)\|z^{k+1} - z\|^2 \\ \leq\ & (1 + 2\tau L + 2|\gamma-\beta| + 4\gamma)\|z^k - z\|^2 + (2\tau L + 2|\gamma-\beta| + 2\gamma)\|z^{k-1} - z\|^2 \\ & + \left(|\gamma-\beta| - \frac{1}{2}\right)\|z^{k+1} - z^{k+0.5}\|^2 + (2\alpha^2L^2 + 2\gamma - 1)\|z^{k+0.5} - z^k\|^2 \\ & +2(\alpha-\eta)\langle F(z^k), z^{k+0.5} - z^{k+1}\rangle, \end{aligned} \tag{2.3.16}$$

then (2.3.15) follows by taking $z = z^*$ and bounding the term $\langle F(z^{k+0.5}), z^{k+0.5} - z^*\rangle$ in the same way as (2.2.10).

We shall follow a similar logic to the proof for extra-gradient method established in Section 2.2.1. The optimality condition for $z^{k+0.5}$ is given by:

$$\langle z^{k+0.5} - (z^k - \eta F(z^k) + \beta(z^k - z^{k-1})), z - z^{k+0.5}\rangle \geq 0, \quad \forall z \in \mathcal{Z},$$

which can be rewritten as the following:

$$\begin{aligned} & \langle \eta F(z^k) - \beta(z^k - z^{k-1}), z - z^{k+0.5}\rangle \\ \geq\ & \frac{1}{2}\left(\|z^{k+0.5} - z^k\|^2 + \|z^{k+0.5} - z\|^2 - \|z - z^k\|^2\right), \quad \forall z \in \mathcal{Z}. \end{aligned} \tag{2.3.17}$$

On the other hand, the optimality condition of z^{k+1} is given by:

$$\langle z^{k+1} - \left(z^k - \alpha F(z^{k+0.5}) + \gamma(z^k - z^{k-1}) - \tau\left[F(z^k) - F(z^{k-1})\right]\right), z - z^{k+1}\rangle \geq 0, \quad \forall z \in \mathcal{Z},$$

which results in

$$\begin{aligned}
& \langle \alpha F(z^{k+0.5}) - \gamma(z^k - z^{k-1}) + \tau \left[F(z^k) - F(z^{k-1}) \right], z - z^{k+1} \rangle \\
\geq \; & \frac{1}{2} \left(\|z^{k+1} - z^k\|^2 + \|z^{k+1} - z\|^2 - \|z - z^k\|^2 \right), \quad \forall z \in \mathcal{Z}.
\end{aligned} \tag{2.3.18}$$

Continuing from (2.3.18), the next inequalities follow:

$$\begin{aligned}
& \frac{1}{2} \left(\|z^{k+1} - z^k\|^2 + \|z^{k+1} - z\|^2 - \|z - z^k\|^2 \right) \\
\leq \; & \langle \alpha F(z^{k+0.5}) - \gamma(z^k - z^{k-1}) + \tau \left[F(z^k) - F(z^{k-1}) \right], z - z^{k+1} \rangle \\
= \; & \alpha \langle F(z^{k+0.5}), z - z^{k+0.5} \rangle + \alpha \langle F(z^{k+0.5}), z^{k+0.5} - z^{k+1} \rangle \\
& - \gamma \langle z^k - z^{k-1}, z - z^{k+1} \rangle + \tau \langle F(z^k) - F(z^{k-1}), z - z^{k+1} \rangle \\
= \; & \alpha \langle F(z^{k+0.5}), z - z^{k+0.5} \rangle - \gamma \langle z^k - z^{k-1}, z - z^{k+1} \rangle + \tau \langle F(z^k) - F(z^{k-1}), z - z^{k+1} \rangle \\
& + \alpha \langle F(z^{k+0.5}) - F(z^k), z^{k+0.5} - z^{k+1} \rangle + (\alpha - \eta) \langle F(z^k), z^{k+0.5} - z^{k+1} \rangle \\
& + \eta \langle F(z^k), z^{k+0.5} - z^{k+1} \rangle.
\end{aligned} \tag{2.3.19}$$

We shall establish the bounds for four terms on the RHS of the above expression (2.3.19). Firstly,

$$\begin{aligned}
& -\gamma \langle z^k - z^{k-1}, z - z^{k+1} \rangle = \gamma \langle z^k - z^{k-1}, z^{k+1} - z^{k+0.5} \rangle + \gamma \langle z^k - z^{k-1}, z^{k+0.5} - z \rangle \\
\leq \; & \gamma \langle z^k - z^{k-1}, z^{k+1} - z^{k+0.5} \rangle + \frac{\gamma}{2} \left(\|z^k - z^{k-1}\|^2 + \|z^{k+0.5} - z\|^2 \right) \\
\leq \; & \gamma \langle z^k - z^{k-1}, z^{k+1} - z^{k+0.5} \rangle + \gamma \left(\|z^{k-1} - z\|^2 + \|z^{k+0.5} - z^k\|^2 + 2\|z^k - z\|^2 \right).
\end{aligned} \tag{2.3.20}$$

Secondly,

$$\begin{aligned}
& \tau \langle F(z^k) - F(z^{k-1}), z - z^{k+1} \rangle \\
\leq \; & \tau \|F(z^k) - F(z^{k-1})\| \|z - z^{k+1}\| \\
\leq \; & \frac{\tau L}{2} \left(\|z^k - z^{k-1}\|^2 + \|z - z^{k+1}\|^2 \right) \\
\leq \; & \frac{\tau L}{2} \left(2\|z^k - z\|^2 + 2\|z^{k-1} - z\|^2 + \|z^{k+1} - z\|^2 \right).
\end{aligned} \tag{2.3.21}$$

Thirdly,

$$\begin{aligned}
& \alpha \langle F(z^{k+0.5}) - F(z^k), z^{k+0.5} - z^{k+1} \rangle \leq \alpha \|F(z^{k+0.5}) - F(z^k)\| \|z^{k+0.5} - z^{k+1}\| \\
\leq \; & \alpha L \|z^{k+0.5} - z^k\| \|z^{k+0.5} - z^{k+1}\| \leq \alpha^2 L^2 \|z^{k+0.5} - z^k\|^2 + \frac{1}{4} \|z^{k+0.5} - z^{k+1}\|^2.
\end{aligned} \tag{2.3.22}$$

Finally, using (2.3.17) with $z = z^{k+1}$:

$$\begin{aligned} & \eta\langle F(z^k), z^{k+0.5} - z^{k+1}\rangle \\ \leq\ & -\beta\langle z^k - z^{k-1}, z^{k+1} - z^{k+0.5}\rangle \\ & +\frac{1}{2}\left(\|z^{k+1} - z^k\|^2 - \|z^{k+0.5} - z^k\|^2 - \|z^{k+1} - z^{k+0.5}\|^2\right). \end{aligned} \tag{2.3.23}$$

Using (2.3.20)-(2.3.23) in inequality (2.3.19), we get the following relation by cancelling and combining terms:

$$\begin{aligned} & \|z^{k+1} - z\|^2 - \|z^k - z\|^2 + 2\alpha\langle F(z^{k+0.5}), z^{k+0.5} - z\rangle \\ \leq\ & 2(\gamma - \beta)\langle z^k - z^{k-1}, z^{k+1} - z^{k+0.5}\rangle + \tau L\|z^{k+1} - z\|^2 \\ & +(2\tau L + 4\gamma)\|z^k - z\|^2 + (2\tau L + 2\gamma)\|z^{k-1} - z\|^2 \\ & +(2\gamma + 2\alpha^2 L^2 - 1)\|z^{k+0.5} - z^k\|^2 - \frac{1}{2}\|z^{k+1} - z^{k+0.5}\|^2 \\ & +2(\alpha - \eta)\langle F(z^k), z^{k+0.5} - z^{k+1}\rangle. \end{aligned}$$

We note the following bound:

$$\begin{aligned} & (\gamma - \beta)\langle z^k - z^{k-1}, z^{k+1} - z^{k+0.5}\rangle \leq \frac{|\gamma - \beta|}{2}\left(\|z^k - z^{k-1}\|^2 + \|z^{k+1} - z^{k+0.5}\|^2\right) \\ \leq\ & \frac{|\gamma - \beta|}{2}\left(2\|z^k - z\|^2 + 2\|z^{k-1} - z\|^2 + \|z^{k+1} - z^{k+0.5}\|^2\right), \end{aligned}$$

then by rearranging the terms we get (2.3.16). □

We make the following observations based on Lemma 2.3.1:

1. To be able to rectify the bounds and relate only to the quantities $\|z^{k+1} - z^*\|^2$, $\|z^k - z^*\|^2$ and $\|z^{k-1} - z^*\|^2$, we need to let the coefficients of the terms $\|z^{k+1} - z^{k+0.5}\|^2$ and $\|z^{k+0.5} - z^k\|^2$ be non-positive, and let the coefficient of $F(z^k)^\top(z^{k+1} - z^{k+0.5})$ be 0.

2. A valid linear reduction requires:

$$0 < \tau L < \alpha\mu - 4\gamma - 2|\gamma - \beta| - 2\tau L < 1.$$

3. The coefficient of $\|z^{k-1} - z^*\|^2$ need be less than the difference between the coefficients of $\|z^{k+1} - z^*\|^2$ and $\|z^k - z^*\|^2$, i.e.

$$2\gamma + 2|\gamma - \beta| + 2\tau L < (\alpha\mu - 4\gamma - 2|\gamma - \beta| - 2\tau L) - \tau L.$$

The above three observations lead the relation in Lemma 2.3.1 to take the form

$$(1-u)\|z^{k+1}-z^*\|^2 \le (1-s)\|z^k-z^*\|^2 + t\|z^{k-1}-z^*\|^2, \quad 0 \le t \le s-u < 1-u. \tag{2.3.24}$$

Dividing both sides by $1-u$, we have

$$\begin{aligned} \|z^{k+1}-z^*\|^2 &\le \left(1-\frac{s-u}{1-u}\right)\|z^k-z^*\|^2 + \frac{t}{1-u}\|z^{k-1}-z^*\|^2 \\ &= (1-a)\|z^k-z^*\|^2 + b\|z^{k-1}-z^*\|^2, \end{aligned} \tag{2.3.25}$$

with $a = \frac{s-u}{1-u}, \quad b = \frac{t}{1-u}$. The above bound can then be further transformed as follows:

Lemma 2.3.2. *Suppose that* (2.3.25) *holds. Then, for any ν satisfying*

$$b < \frac{\sqrt{(1-a)^2+4b}-(1-a)}{2} \le \nu < a, \tag{2.3.26}$$

it holds that

$$\|z^{k+1}-z^*\|^2 + \nu\|z^k-z^*\|^2 \le (1-(a-\nu))\|z^k-z^*\|^2 + \nu(1-(a-\nu))\|z^{k-1}-z^*\|^2.$$

Proof. The proof follows immediately from the fact that $b \le \nu(1-(a-\nu))$. □

A specific choice of ν can be $\frac{a+b}{2}$. This leads to our main convergence result, summarized in the next theorem.

Theorem 2.3.3. *For solving VI model* (2.1.1)*, with the mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ being Lipschitz continuous with constant L and strongly monotone with constant μ, the sequence $\{z^k\}$, $k = 0, 1, 2, \ldots$ generated from the scheme* (2.3.14) *yields a linear convergence as the following:*

$$\|z^{k+1}-z^*\|^2 + \nu\|z^k-z^*\|^2 \le (1-(a-\nu))\|z^k-z^*\|^2 + \nu(1-(a-\nu))\|z^{k-1}-z^*\|^2,$$

where

$$\begin{aligned} &a = \frac{\alpha\mu - 4\gamma - 2|\gamma-\beta| - 3\tau L}{1-\tau L}, \quad b = \frac{2\gamma + 2|\gamma-\beta| + 2\tau L}{1-\tau L}, \\ &\frac{\sqrt{(1-a)^2+4b}-(1-a)}{2} \le \nu < a, \end{aligned} \tag{2.3.27}$$

provided that the parameters $\alpha, \beta, \gamma, \eta, \tau$ in (2.3.14) *satisfy*

$$\begin{cases} 0 < \alpha\mu - 6\gamma - 4|\gamma-\beta| - 5\tau L, \\ |\gamma-\beta| \le \frac{1}{2}, \\ 2\alpha^2L^2 + 2\alpha\mu + 2\gamma \le 1, \\ \eta = \alpha, \\ \alpha, \beta, \gamma, \eta, \tau \ge 0. \end{cases} \tag{2.3.28}$$

Proof. The proof follows directly from Lemma 2.3.1 and Lemma 2.3.2. Note that we omit the constraint suggested by the second observation following Lemma 2.3.1, which is redundant given the first and third constraint in (2.3.28). □

As an example, we may choose $\alpha = \eta = \frac{1}{4L}$, $\beta = \gamma = \frac{\mu}{64L}$, and $\tau = \frac{\mu}{64L^2}$ to satisfy (2.3.28). Then,

$$\left(1 - \frac{\mu}{64L}\right) \|z^{k+1} - z^*\|^2 \leq \left(1 - \frac{5\mu}{32L}\right) \|z^k - z^*\|^2 + \frac{\mu}{16L} \|z^{k-1} - z^*\|^2.$$

Denote $\frac{\mu}{L} = \sigma$ and divide both sides with $1 - \frac{\sigma}{64}$, we get:

$$\begin{aligned} \|z^{k+1} - z^*\|^2 &\leq \left(1 - \tfrac{\frac{9\sigma}{64}}{1-\frac{\sigma}{64}}\right) \|z^k - z^*\|^2 + \tfrac{\frac{\sigma}{16}}{1-\frac{\sigma}{64}} \|z^{k-1} - z^*\|^2 \\ &= \left(1 - \tfrac{9\sigma}{64-\sigma}\right) \|z^k - z^*\|^2 + \tfrac{\frac{\sigma}{16}}{1-\frac{\sigma}{64}} \|z^{k-1} - z^*\|^2 \\ &\leq \left(1 - \tfrac{9\sigma}{64}\right) \|z^k - z^*\|^2 + \tfrac{4\sigma}{63} \|z^{k-1} - z^*\|^2. \end{aligned}$$

By fixing $\nu = \frac{a+b}{2}$, and with such choices of parameters, the reduction rate is guaranteed to be at least $1 - \frac{1}{2} \cdot \left(\frac{9\sigma}{64} - \frac{4\sigma}{63}\right) < 1 - \frac{\sigma}{32}$. The resulting iteration complexity therefore reaches the *optimal* order of magnitude $\mathcal{O}\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$, for finding an ϵ-solution.

Since the scheme (2.3.14) follows the extra-point approach where the lower bound takes effect [108], this iteration complexity is indeed optimal, the same as OGDA or extra-gradient method. However, a general scheme such as (2.3.14) allows trainable parameter-tuning to achieve practical high performance, while maintaining a theoretical optimal iteration complexity bound. As we shall see in Section 2.6, given the problem at hand, simpler configurations such as those in Table 2.1 may not always perform the best and can be outperformed by some non-trivial parameter combinations. This necessitates the flexibility allowed in the extra-point approach.

2.4 An Extra-Point Approach to a Class of Strongly Monotone VI Problems

In this section, we consider another exemplifying realization of the extra-point approach in an extended class of VI. The operator $F(z)$ in this class of VI can be explicitly expressed in the sum of a general vector mapping $H(z)$ and a gradient mapping $\nabla g(z)$ from a convex function $g(z)$:

$$F(z) := H(z) + \nabla g(z), \tag{2.4.29}$$

where we assume $H : \mathbb{R}^n \mapsto \mathbb{R}^n$ is strongly monotone with modulus $\mu_h > 0$ and Lipschitz continuous with constant L_h in $\mathcal{Z}$ and $\nabla g(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^n$ is monotone with modulus $\mu_g \geq 0$

and Lipschitz continuous with constant L_g in $\mathcal{Z}$. The presence of the explicit gradient mapping $\nabla g(\cdot)$ provides a more accurate estimation on the continuity of the overall operator $F(\cdot)$, which results in a combined improved iteration complexity as we shall see in the later analysis.

The original work of analyzing iteration complexity of the type of VI (2.4.29) through an "accelerated mirror-prox method" is established in [10], where the stochastic setting is also considered with the operators $H(\cdot)$ and $\nabla g(\cdot)$ being only monotone. They show that the overall iteration complexity of the proposed method is

$$\frac{L_h}{\epsilon} + \sqrt{\frac{L_g}{\epsilon}}, \tag{2.4.30}$$

which is optimal for the type of VI (the stochastic component in the result is omitted). The method combines the mirror-prox method [62] for monotone VI, $H(\cdot)$, with two extra sequences similar to the Nesterov-type acceleration for convex $g(\cdot)$. We notice that the accelerated mirror-prox method is also a specific realization of the extra-point approach. However, as shown in [10], the specialized form is indeed more involved in order to exploit the special problem structure, i.e. the gradient mapping. In this section, we focus on the case when $H(\cdot)$ is strongly monotone instead and propose a scheme of extra points that is similar to the accelerated mirror-prox method in its form but include other search directions as proposed in (2.3.14). We then show its accelerated iteration complexity compared to general strongly monotone VI.

The proposed extra-point scheme updates the iterate as follows:

(2.4.31)
$$\begin{cases} y^k &= (1-\theta)v^k + \theta z^k, \\ z^{k+0.5} &= P_{\mathcal{Z}}\left(z^k + \beta(z^k - z^{k-1}) - \eta\left(H(z^k) + \nabla g(y^k)\right)\right), \\ z^{k+1} &= P_{\mathcal{Z}}\left(z^k - \alpha\left(H(z^{k+0.5}) + \nabla g(y^k)\right) + \gamma(z^k - z^{k-1}) - \tau\left(H(z^k) - H(z^{k-1})\right)\right), \\ v^{k+1} &= (1-\theta)v^k + \theta z^{k+0.5} \end{cases}$$

for $k = 0, 1, 2, ...$ and $z^{-1} := z^0 = v^0$, with $\alpha, \beta, \gamma, \eta, \tau \geq 0$ and $\theta \in (0, 1)$.

In the above scheme, two extra-point sequences $\{y^k\}$ and $\{v^k\}$ are introduced in addition to $\{z^{k+0.5}\}$, compared to the previous scheme (2.3.14) proposed for general strongly monotone VI. The purpose of these sequences is to make explicit use of the improved estimation on the continuity of the gradient mapping $\nabla g(\cdot)$ in terms of L_g. The analysis consists of three parts: the first part analyzes the extra-point sequence $\{z^{k+0.5}\}$ and the main iterate sequence $\{z^k\}$, where the derivations follow similar logic for analyzing general strongly

monotone VI with operator $H(\cdot) + \nabla g(y^k)$ at each iteration; the second part analyzes the two extra-point sequences $\{y^k\}$ and $\{v^k\}$, which helps establish relations in function value $g(\cdot)$. The final part combines the previous results to determine suitable parameters that lead to accelerated iteration complexity.

To proceed with the first part of the analysis, let us define a new mapping:

$$\hat{F}(z) := H(z) + \nabla g(y^k).$$

By adding a constant mapping $\nabla g(y^k)$, $\hat{F}(\cdot)$ retains the same properties as $H(\cdot)$ (strong monotonicity, Lipschitz continuity). This enables us to derive the next inequality following the similar logic in the proof of Lemma 2.3.1 by replacing $F(z)$ with $\hat{F}(z)$, which holds for all $z \in \mathcal{Z}$:

$$\begin{aligned}
& 2\alpha\langle H(z^{k+0.5}) + \nabla g(y^k), z^{k+0.5} - z\rangle + (1 - \tau L_h)\|z^{k+1} - z\|^2 \\
\leq\ & (1 + 2\tau L_h + 2|\gamma - \beta| + 4\gamma)\|z^k - z\|^2 + (2\tau L_h + 2|\gamma - \beta| + 2\gamma)\|z^{k-1} - z\|^2 \\
& + \left(|\gamma - \beta| - \frac{1}{2}\right)\|z^{k+1} - z^{k+0.5}\|^2 + (2\alpha^2 L_h^2 + 2\gamma - 1)\|z^{k+0.5} - z^k\|^2 \\
& +2(\alpha - \eta)\langle H(z^k) + \nabla g(y^k), z^{k+0.5} - z^{k+1}\rangle. \qquad (2.4.32)
\end{aligned}$$

Note that (2.4.32) can be viewed as the inequality preceding (2.3.15), where we have expressed $\hat{F}(z)$ explicitly. With the strong monotonicity of $H(\cdot)$,

$$\begin{aligned}
& \langle H(z^{k+0.5}) + \nabla g(y^k), z^{k+0.5} - z\rangle \\
\geq\ & \langle H(z) + \nabla g(y^k), z^{k+0.5} - z\rangle + \mu_h\|z^{k+0.5} - z\|^2 \\
\geq\ & \langle H(z) + \nabla g(y^k), z^{k+0.5} - z\rangle + \frac{\mu_h}{2}\|z^k - z\|^2 - \mu_h\|z^{k+0.5} - z^k\|^2.
\end{aligned}$$

Then by introducing the following constraints:

$$\alpha = \eta, \quad |\gamma - \beta| \leq \frac{1}{2}, \qquad (2.4.33)$$

together with changes of notations:

$$u = \tau L_h, \quad s = \alpha\mu_h - 2\tau L_h - 2|\gamma - \beta| - 4\gamma, \quad t = 2(\tau L_h + |\gamma - \beta| + \gamma),$$

we can simplify (2.4.32) into:

$$\begin{aligned}
& 2\alpha\langle H(z) + \nabla g(y^k), z^{k+0.5} - z\rangle + (1 - u)\|z^{k+1} - z\|^2 \\
\leq\ & (1 - s)\|z^k - z\|^2 + t\|z^{k-1} - z\|^2 + (2\alpha\mu_h + 2\alpha^2 L_h^2 + 2\gamma - 1)\|x^{k+0.5} - x^k\|^2, \qquad (2.4.34)
\end{aligned}$$

which concludes the first part of the analysis.

In the second part, we shall analyze the effect of $\{y^k\}$ and $\{v^k\}$ on the function $g(\cdot)$. This procedure follows a similar logic to the analysis in [10]. With the Lipschitz continuity of $g(\cdot)$,

$$\begin{aligned}
g(v^{k+1}) &\leq g(y^k) + \langle \nabla g(y^k), v^{k+1} - y^k \rangle + \frac{L_g}{2}\|v^{k+1} - y^k\|^2 \\
&= g(y^k) + \langle \nabla g(y^k), (1-\theta)v^k + \theta x^{k+0.5} - y^k \rangle + \frac{L_g\theta^2}{2}\|z^{k+0.5} - z^k\|^2 \\
&= (1-\theta)\Big(g(y^k) + \langle \nabla g(y^k), v^k - y^k \rangle\Big) \\
&\quad +\theta\Big(g(y^k) + \langle \nabla g(y^k), z^{k+0.5} - y^k \rangle\Big) + \frac{L_g\theta^2}{2}\|z^{k+0.5} - z^k\|^2 \\
&\leq (1-\theta)g(v^k) + \theta g(z) + \theta\langle \nabla g(y^k), z^{k+0.5} - z \rangle + \frac{L_g\theta^2}{2}\|z^{k+0.5} - z^k\|^2, \qquad (2.4.35)
\end{aligned}$$

which holds for all $z \in \mathcal{Z}$.

We are now ready to combine the results from the previous parts and derive the final convergence results. Denote

$$Q_k(z) := \langle H(z), v^k - z \rangle + g(v^k) - g(z) \qquad (2.4.36)$$

as a function of $z \in \mathbb{R}^n$ at a fixed v^k. We have the following:

$$\begin{aligned}
Q_{k+1}(z) &= \langle H(z), v^{k+1} - z \rangle + g(v^{k+1}) - g(z) \\
&= (1-\theta)\langle H(z), v^k - z \rangle + \theta\langle H(z), z^{k+0.5} - z \rangle + g(v^{k+1}) - g(z) \\
&\overset{(2.4.35)}{\leq} (1-\theta)\langle H(z), v^k - z \rangle + \theta\langle H(z), z^{k+0.5} - z \rangle \\
&\quad +(1-\theta)\Big(g(v^k) - g(z)\Big) + \theta\langle \nabla g(y^k), z^{k+0.5} - z \rangle + \frac{L_g\theta^2}{2}\|z^{k+0.5} - z^k\|^2 \\
&= (1-\theta)\Big[\langle H(z), v^k - z \rangle + g(v^k) - g(z)\Big] \\
&\quad +\theta\left[\langle H(z) + \nabla g(y^k), z^{k+0.5} - z \rangle + \frac{L_g\theta}{2}\|z^{k+0.5} - z^k\|^2\right] \\
&\overset{(2.4.34)}{\leq} (1-\theta)Q_k(z) \\
&\quad +\frac{\theta}{2\alpha}\Big((1-s)\|z^k - z\|^2 - (1-u)\|z^{k+1} - z\|^2 + t\|z^{k-1} - z\|^2\Big) \\
&\quad -\frac{\theta}{2\alpha}(1 - 2\alpha\mu_h - 2\alpha^2 L_h^2 - 2\gamma - L_g\theta\alpha)\|z^{k+0.5} - z^k\|^2.
\end{aligned}$$

Rearranging the terms in the above inequality gives us:

$$\begin{aligned}
& Q_{k+1}(z) + \frac{\theta}{2\alpha}(1-u)\|z^{k+1} - z\|^2 \\
\leq\ & (1-\theta)Q_k(z) + \frac{\theta}{2\alpha}\Big((1-s)\|z^k - z\|^2 + t\|z^{k-1} - z\|^2\Big) \\
& -\frac{\theta}{2\alpha}(1 - 2\alpha\mu_h - 2\alpha^2 L_h^2 - 2\gamma - L_g\theta\alpha)\|z^{k+0.5} - z^k\|^2.
\end{aligned}$$

We observe that the above inequality resembles the intermediate result derived in the previous section (cf. (2.3.24)) but with the presence of the term $Q_k(z)$, which stems from the extra-point sequences $\{y^k\}, \{v^k\}$ and the resulting bound on the function value $g(\cdot)$. It is then combined with the bound (2.4.34) derived from the sequences $\{z^k\}, \{z^{k+0.5}\}$. We shall push one step forward for the bound on $\|z^k - z\|^2$ by applying the same argument as in (2.3.25) and Lemma 2.3.2. The following constraints in addition to (2.4.33) are in place:

$$\begin{cases} 2\alpha\mu_h + 2\alpha^2 L_h^2 + 2\gamma + L_g\theta\alpha \leq 1, \\ \\ t \leq s - u < 1 - u. \end{cases} \tag{2.4.37}$$

Denote $a = \frac{s-u}{1-u}$ and $b = \frac{t}{1-u}$, Lemma 2.3.2 states that the following inequality holds with a constant ν such that (2.3.26) is satisfied (e.g. $\nu = \frac{a+b}{2}$):

$$\begin{aligned} & Q_{k+1}(z) + \frac{\theta}{2\alpha}(1-u)\left(\|z^{k+1} - z\|^2 + \nu\|z^k - z\|^2\right) \\ \leq \quad & (1-\theta)Q_k(z) + \frac{\theta}{2\alpha}(1-u)(1-(a-\nu))\left(\|z^k - z\|^2 + \nu\|z^{k-1} - z\|^2\right), \end{aligned}$$

which by further requiring

$$\theta \leq a - \nu \tag{2.4.38}$$

gives us

$$\begin{aligned} & Q_{k+1}(z) + \frac{\theta}{2\alpha}(1-u)\left(\|z^{k+1} - z\|^2 + \nu\|z^k - z\|^2\right) \\ \leq \quad & (1-\theta)\left[Q_k(z) + \frac{\theta}{2\alpha}(1-u)\left(\|z^k - z\|^2 + \nu\|z^{k-1} - z\|^2\right)\right], \\ \leq \quad & (1-\theta)^{k+1}\left[Q_0(z) + \frac{\theta}{2\alpha}(1-u)\left(\|z^0 - z\|^2 + \nu\|z^{-1} - z\|^2\right)\right], \\ = \quad & (1-\theta)^{k+1}\left[Q_0(z) + \frac{\theta}{\alpha}(1-u)(1+\nu)\|z^0 - z\|^2\right]. \end{aligned} \tag{2.4.39}$$

Note that $z^{-1} := z^0 = v^0$. The above inequality holds for all $z \in \mathcal{Z}$, regardless of the sign of $Q_k(z) + \frac{\theta}{2\alpha}(1-u)\left(\|z^k - z\|^2 + \nu\|z^{k-1} - z\|^2\right)$. The next lemma shows that $\max_{z\in\mathcal{Z}} Q_k(z)$ is a valid merit function with respect to the VI with the combined mapping (2.4.29). Therefore the convergence of $\max_{z\in\mathcal{Z}} Q_k(z)$ guarantees us an approximate solution v^k.

Lemma 2.4.1. *For $\{v^k\}$ generated by (2.4.31) and $Q_k(z)$ defined as (2.4.36), we have $\max_{z\in\mathcal{Z}} Q_k(z) \geq 0$ and $\max_{z\in\mathcal{Z}} Q_k(z) = 0$ if and only if v^k solves the VI with $F(z) := H(z) + \nabla g(z)$.*

Proof. It is clear that $\max_{z\in\mathcal{Z}} Q_k(z) \geq 0$, and we shall show the second statement. If $\max_{z\in\mathcal{Z}} Q_k(z) = 0$, we have

$$\begin{aligned} 0 = \max_{z\in\mathcal{Z}} Q_k(z) &= \max_{z\in\mathcal{Z}} \quad \langle H(z), v^k - z\rangle + g(v^k) - g(z) \\ &\geq \max_{z\in\mathcal{Z}} \quad \langle H(z) + \nabla g(z), v^k - z\rangle, \end{aligned}$$

where the inequality is due to the convexity of $g(\cdot)$. Therefore, we have $\langle H(z) + \nabla g(z), z - v^k\rangle \geq 0$ for all $z \in \mathcal{Z}$, which implies that v^k is a weak solution to the VI with $F(z) := H(z) + \nabla g(z)$. Since we assume $F(\cdot)$ to be continuous, it is also a strong solution, i.e.

$$\langle H(v^k) + \nabla g(v^k), z - v^k\rangle \geq 0, \quad \forall z \in \mathcal{Z}.$$

On the other hand, if v^k is a (strong) solution to the VI, then

$$\begin{aligned} Q_k(z) &= \langle H(z), v^k - z\rangle + g(v^k) - g(z) \leq \langle H(v^k), v^k - z\rangle + \langle \nabla g(v^k), v^k - z\rangle \\ &= \langle H(v^k) + \nabla g(v^k), v^k - z\rangle \leq 0, \quad \forall z \in \mathcal{Z}. \end{aligned}$$

Therefore, $\max_{z\in\mathcal{Z}} Q_k(z) \leq 0$, which together with the first statement implies $\max_{z\in\mathcal{Z}} Q_k(z) = 0$. □

Collecting the necessary constraints on the parameters (2.4.33),(2.4.37),(2.4.38), we are now ready to summarize the convergence result of the proposed scheme (2.4.31) in the following theorem.

Theorem 2.4.2. *For the VI with constraint $\mathcal{Z}$ and $F(z) := H(z) + \nabla g(z)$, the iterates generated by the scheme (2.4.31) for $k = 0, 1, 2, \ldots$ satisfy the following inequality with $Q_k(z)$ defined in (2.4.36):*

$$\max_{z\in\mathcal{Z}} Q_k(z) \leq (1-\theta)^k \max_{z\in\mathcal{Z}} \left[Q_0(z) + \frac{\theta}{\alpha}(1 - \tau L_h)(1+\nu)\|z^0 - z\|^2 \right],$$

given the following constraints on the parameters $\alpha, \beta, \gamma, \eta, \tau \geq 0$ and $\theta \in (0,1)$ are satisfied:

$$\begin{cases} \alpha = \eta, \quad |\gamma - \beta| \leq \frac{1}{2} \\ 2\alpha\mu_h + 2\alpha^2 L_h^2 + 2\gamma + L_g\theta\alpha \leq 1 \\ 5\tau L_h + 4|\gamma - \beta| + 6\gamma \leq \alpha\mu_h \\ \theta + \nu \leq a, \quad b < \frac{\sqrt{(1-a)^2 + 4b} - (1-a)}{2} \leq \nu < a \end{cases} \tag{2.4.40}$$

where

$$a = \frac{\alpha\mu_h - 4\gamma - 2|\gamma - \beta| - 3\tau L_h}{1 - \tau L_h}, \quad b = \frac{2\gamma + 2|\gamma - \beta| + 2\tau L_h}{1 - \tau L_h}.$$

Finally, we shall provide a specific example of parameter choices that satisfy the constraint (2.4.40) and demonstrate an explicit iteration complexity. Consider

$$(\alpha,\beta,\gamma,\eta,\tau) = \left(\tfrac{1}{4\tilde{L}}, \tfrac{\mu_h}{64\tilde{L}}, \tfrac{\mu_h}{64\tilde{L}}, \tfrac{1}{4\tilde{L}}, \tfrac{\mu_h}{64\tilde{L}L_h}\right), \quad \tilde{L} := L_h + \sqrt{L_g\mu_h}$$
$$\theta = \tfrac{1}{64}\min\left(\sqrt{\tfrac{\mu_h}{L_g}}, \tfrac{\mu_h}{L_h}\right), \qquad \nu = \frac{a+b}{2}.$$

While most of the constraints are clearly satisfied, we only verify the ones involving ν. Note that we have $a+b \geq \sqrt{(1-a)^2+4b} - (1-a)$ and

$$a = \frac{\frac{9\mu_h}{64\tilde{L}}}{1-\frac{\mu_h}{64\tilde{L}}}, \quad b = \frac{\frac{4\mu_h}{64\tilde{L}}}{1-\frac{\mu_h}{64\tilde{L}}}, \quad a-\nu = \frac{a-b}{2} = \frac{1}{2}\cdot\frac{\frac{5\mu_h}{64\tilde{L}}}{1-\frac{\mu_h}{64\tilde{L}}} \geq \frac{5\mu_h}{128\tilde{L}}.$$

Since $\tilde{L} \leq 2\max\left(L_h, \sqrt{L_g\mu_h}\right)$, we have $a-\nu \geq \theta$. To determine the iteration complexity, note that

$$\frac{\theta}{\alpha} = \frac{1}{16}\min\left(\sqrt{\frac{\mu_h}{L_g}}, \frac{\mu_h}{L_h}\right)\cdot(L_h+\sqrt{L_g\mu_h}) \leq \frac{\mu_h}{8},$$

and $(1-\tau L_h)(1+\nu) = 1 + \frac{11\mu_h}{128\tilde{L}}$. Therefore,

$$\begin{aligned}
\max_{z\in\mathcal{Z}} Q_k(z) &\leq (1-\theta)^k \max_{z\in\mathcal{Z}}\left[Q_0(z) + \frac{\theta}{\alpha}(1-\tau L_h)(1+\nu)\|z^0-z\|^2\right] \\
&\leq (1-\theta)^k \max_{z\in\mathcal{Z}}\left[Q_0(z) + \frac{\mu_h}{2}\|z^0-z\|^2\right] \\
&= (1-\theta)^k \max_{z\in\mathcal{Z}}\left[\langle H(z), z^0-z\rangle + g(z^0) - g(z) + \frac{\mu_h}{2}\|z^0-z\|^2\right] \\
&\leq (1-\theta)^k \max_{z\in\mathcal{Z}}\left[\langle H(z^0)+\nabla g(z^0), z^0-z\rangle - \frac{\mu_h}{2}\|z^0-z\|^2\right] \\
&\leq (1-\theta)^k \cdot \frac{1}{2\mu_h}\|H(z^0)+\nabla g(z^0)\|^2,
\end{aligned}$$

where in the third inequality we use the strong monotonicity of $H(\cdot)$ and the convexity of $g(\cdot)$. The iteration complexity is then given by

$$\mathcal{O}\left(\frac{1}{\theta}\ln\frac{1}{\epsilon}\right) = \mathcal{O}\left(\max\left(\frac{L_h}{\mu_h}, \sqrt{\frac{L_g}{\mu_h}}\right)\ln\frac{1}{\epsilon}\right). \tag{2.4.41}$$

The scheme (2.4.31) demonstrates a more complicated realization of the extra-point approach by introducing multiple extra-point sequences, and the result shows that the iteration complexity can be improved from $\mathcal{O}\left(\frac{L_h+L_g}{\mu_h}\ln\frac{1}{\epsilon}\right)$ (through an optimal method for solving general strongly monotone VI, e.g. extra-gradient method) to (2.4.41) for solving VI of the form (2.4.29). While the scheme (2.4.31) is generalized from the procedure proposed in [10], the authors in [10] considered the monotone case. Alternatively, it is possible to adopt a *restart* procedure to transform the results for monotone VI to strongly monotone VI (cf. [73]). As we have emphasized in the previous section, the presence of other first-order search directions in the scheme provides the flexibility in searching configurations leading to better practical performance, as we shall see in Section 2.6.

2.5 An Extra-Point Approach to Strongly Convex Optimization

In this section, we consider another realization of the extra-point approach in the context of convex optimization. Note that the analysis in the previous section can not be directly specialized to optimization by taking $H(z) = 0$, since it relies on the strong monotonicity of $H(\cdot)$. Moreover, strongly convex optimization is such an important and specially structured VI class that it deserves a separate treatment. Indeed, it requires a quite different analysis in optimization in order to match the improved lower bound $\mathcal{O}\left(\sqrt{\kappa}\ln\frac{1}{\epsilon}\right)$, as compared to $\mathcal{O}\left(\kappa\ln\frac{1}{\epsilon}\right)$ for a general strongly monotone VI. In particular, the proposed scheme uses Nesterov's accelerated method for optimization as its basic structure, combined with additional extra points and search directions as we have introduced before. Since Nesterov's accelerated method pioneers most other known accelerated method, it makes sense for our exemplifying scheme to adopt a similar form but at the same time, take advantage of the flexibility allowed by the extra-point approach. We start with introducing some more background on accelerated methods for convex optimization in Section 2.5.1 given the rich developments in the field. A quick review of the analysis for vanilla gradient descent method is also presented, which motivates the use of extra points in the proposed scheme that is formally introduced and analyzed in Section 2.5.2.

2.5.1 Background and preparations

One of the first methods that utilizes the concept of "momentum" may be traced back to the heavy-ball method (2.2.12) proposed by Polyak in [78]. However, the accelerated convergence rate $1-\sqrt{\sigma}$ is only valid for a quadratic objective function, instead of general strongly convex function. The first accelerated method for general smooth convex optimization is Nesterov's method proposed in [65], which admits an iteration complexity $\mathcal{O}(1/\sqrt{\epsilon})$. In [66], Nesterov introduced a variant of the method to solve strongly convex optimization, with a $1-\sqrt{\sigma}$ linear convergence rate. Among its many equivalent forms, perhaps one of the most well-known form of Nesterov's method is the following updating formula

$$\left\{\begin{array}{rcl} y^k & = & \frac{1}{1+\sqrt{\sigma}}x^k + \frac{\sqrt{\sigma}}{1+\sqrt{\sigma}}v^k, \\ x^{k+1} & = & y^k - \frac{1}{L}\nabla f(y^k), \\ v^{k+1} & = & (1-\sqrt{\sigma})v^k + \sqrt{\sigma}(y^k - \frac{1}{\mu}\nabla f(y^k)). \end{array}\right. \tag{2.5.42}$$

Choosing initial points appropriately, the above can be further reduced to the following updating rule:

$$\left\{\begin{array}{rcl} y^k & = & x^k + \frac{1-\sqrt{\sigma}}{1+\sqrt{\sigma}}(x^k - x^{k-1}), \\ x^{k+1} & = & y^k - \frac{1}{L}\nabla f(y^k), \end{array}\right. \tag{2.5.43}$$

which is the form that we used in Section 2.2. Both formulas (2.5.42) (2.5.43) are in fact examples of using extra-point sequences $\{v^k\}$ and/or $\{y^k\}$ to help update the main iterate sequence $\{x^k\}$.

In recent years, many other accelerated first-order methods have been proposed to achieve the $1-\sqrt{\sigma}$ convergence rate; some of them even have a better constant subsumed in the big O notation, compared to Nesterov's method. For example, the geometric descent method proposed in Bubeck *et al.* [7] updates some ball that contains the optimal solution x^* throughout the iterations while trying to reduce its radius. The quadratic averaging method of Drusvyatskiy *et al.* [16] attempts to maximize the minimum of the convex combination of two quadratic lower bounds at each iteration, which actually produces the same iterative sequence as the geometric descent method. The triple momentum method of Van Scoy *et al.* [98] has the following general form:

$$\left\{\begin{array}{rcl} x^{k+1} & = & (1+\beta)x^k - \beta x^{k-1} - \alpha\nabla f(y^k), \\ y^k & = & (1+\gamma)x^k - \gamma x^{k-1}, \\ v^k & = & (1+\delta)x^k - \delta x^{k-1}. \end{array}\right.$$

Although the method itself uses a different parameter choice of $\alpha, \beta, \gamma, \delta$, such general form can include heavy-ball method ($\gamma = \delta = 0$) and Nesterov's fixed step size method ($\beta = \gamma$, $\delta = 0$) as special cases. The information-theoretic exact method (ITEM) is recently proposed by Taylor and Drori [93], which has a similar form as Nesterov's method:

$$\left\{\begin{array}{rcl} y^k & = & (1-\tau_k)x^k + \tau_k v^k, \\ x^{k+1} & = & y^k - \frac{1}{L}\nabla f(y^k), \\ v^{k+1} & = & (1-\sigma\delta_k)v^k + \delta_k(\sigma y^k - \frac{1}{L}\nabla f(y^k)). \end{array}\right.$$

However, different choices of the parameters δ_k, τ_k and different potential function are used in the analysis. The ITEM achieves the exact lower bound in terms of the measurement $\frac{\|x^N - x^*\|^2}{\|x^0 - x^*\|^2}$ for a given iteration number N (see Drori and Taylor [15]). The method reduces to the triple momentum method by taking the parameters δ_k, τ_k the values of their limits, and it reduces to the optimized gradient method of Kim and Fessler [40] by taking $\mu = 0$. The results in Karimi and Vavasis [37] establish a unified analysis for the conjugate gradient method and Nesterov's accelerated method, showing that the progress of both algorithms can be measured by the decrease of potential functions of the same form. A follow-up

work by the same authors in [38] further includes the geometric descent in the analysis. In the recent monograph by d'Aspremont *et al.* [12], an extensive survey is provided on the accelerated methods, including both convex and strongly convex optimization.

Consider the following optimization model:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{2.5.44}$$

where $f(x)$ is strongly convex with modulus μ and $\nabla f(x)$ is Lipschitz continuous with constant L. Therefore,

$$f(x)+\nabla f(x)^\top(y-x)+\frac{\mu}{2}\|y-x\|^2 \le f(y) \le f(x)+\nabla f(x)^\top(y-x)+\frac{L}{2}\|y-x\|^2,\ \ \forall x,y.$$

Recall that $\sigma=\frac{\mu}{L}$ and $\kappa=\frac{L}{\mu}$. The vanilla gradient descent method with fixed step size has the following update:

$$x^{k+1}=x^k-\frac{1}{L}\nabla f(x^k).$$

We have

$$\begin{array}{rcl} f(x^{k+1}) & \le & f(x^k)+\nabla f(x^k)^\top(x^{k+1}-x^k)+\frac{L}{2}\|x^{k+1}-x^k\|^2 \\ & = & f(x^k)-\frac{1}{L}\|\nabla f(x^k)\|^2+\frac{1}{2L}\|\nabla f(x^k)\|^2 \\ & = & f(x^k)-\frac{1}{2L}\|\nabla f(x^k)\|^2 \\ & \le & f(x)-\nabla f(x^k)^\top(x-x^k)-\frac{\mu}{2}\|x-x^k\|^2-\frac{1}{2L}\|\nabla f(x^k)\|^2, \end{array}$$

for any x.

Substitute x^k and x^* for x in the above inequality, we get:

$$\left\{\begin{array}{l} f(x^{k+1})\le f(x^k)-\frac{1}{2L}\|\nabla f(x^k)\|^2, \\ f(x^{k+1})\le f(x^*)-\nabla f(x^k)^\top(x^*-x^k)-\frac{\mu}{2}\|x^*-x^k\|^2-\frac{1}{2L}\|\nabla f(x^k)\|^2. \end{array}\right.$$

Summing up the two inequalities with the first one multiplied by $1-\theta$ and the second one multiplied by θ, we get:

(2.5.45)

$$\begin{array}{rl} & f(x^{k+1})-f(x^*) \\ \le & (1-\theta)(f(x^k)-f(x^*))+\theta\nabla f(x^k)^\top(x^k-x^*)-\frac{\mu\theta}{2}\|x^k-x^*\|^2-\frac{1}{2L}\|\nabla f(x^k)\|^2 \\ \le & (1-\theta)(f(x^k)-f(x^*))+\frac{\theta}{2\mu}\|\nabla f(x^k)\|^2-\frac{1}{2L}\|\nabla f(x^k)\|^2 \\ \overset{\theta\le\frac{\mu}{L}}{\le} & (1-\theta)(f(x^k)-f(x^*)). \end{array}$$

In the second inequality, the following bound is used:

$$\theta\nabla f(x^k)^\top(x^k-x^*)-\frac{\mu\theta}{2}\|x^k-x^*\|^2\le\frac{\theta}{2\mu}\|\nabla f(x^k)\|^2.$$

Since $\theta\le\frac{\mu}{L}=\sigma$, the per-iteration convergence rate is $1-\sigma$, giving an iteration complexity $\kappa\ln(1/\epsilon)$ in order to get $f(x^k)-f(x^*)\le\epsilon$, i.e. an ϵ-solution.

2.5.2 An extra-point scheme

As shown in (2.5.45), directly bounding the term

$$\theta \nabla f(x^k)^\top (x^k - x^*) - \frac{\mu\theta}{2}\|x^k - x^*\|^2 - \frac{1}{2L}\|\nabla f(x^k)\|^2 \tag{2.5.46}$$

requires θ to be small enough, in particular, in the order of σ. This restricts the convergence rate to be $1-\sigma$ instead of the optimal $1-\sqrt{\sigma}$. This explains the limitation of using a single main iterate sequence to guide the algorithm, which motivates introducing extra points and additional search directions to help the update of the main iterates. In addition, instead of just taking $f(x^k) - f(x^*)$ as our measurement of the progress of the algorithm, we shall construct a potential function from an extra sequence of points $\{v^k\}$:

$$f(x^k) - f(x^*) + C\|v^k - x^*\|^2$$

for some constant C, specifically to help cancel out the terms similar to those in (2.5.46) instead of directly bounding them.

Let us now consider the following iterative procedure

$$\left\{\begin{array}{lcl} p^k & := & \nu_1 x^k + \nu_2 v^k \\ y^k & \text{satisfies} & \nabla f(y^k)^\top (p^k - y^k) \le 0 \\ z^k & := & y^k - \frac{\eta}{L}\nabla f(y^k) \\ x^{k+1} & := & y^k - \frac{\alpha}{L}\nabla f(z^k) - \frac{\tau}{L}\left(\nabla f(z^k) - \nabla f(y^k)\right) + \beta(z^k - y^k) \\ v^{k+1} & := & \omega_1 v^k + \omega_2 y^k - \omega_3 \nabla f(y^k), \end{array}\right. \tag{2.5.47}$$

where ν_1, ν_2, $\alpha, \beta, \eta, \tau$, and $\omega_1, \omega_2, \omega_3$ are some positive parameters. As examples, y^k may be taken simply as p^k, or perhaps $y^k := p^k - \frac{1}{L}\nabla f(p^k)$ among other choices. While the main structure in the scheme (2.5.47) and the following analysis resemble the logic in Nesterov's accelerated method, it also introduces additional extra-point sequences $\{y^k\}$ and $\{z^k\}$ as well as other first-order search directions, as we have seen in the previous sections, so as to further take advantage of the flexibility under the general extra-point approach.

To see how the extra points work in procedure (2.5.47), consider the following analysis. First, we have:

$$\begin{aligned} f(x^{k+1}) &\le f(y^k) + \nabla f(y^k)^\top (x^{k+1} - y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 \\ &\le f(x) - \nabla f(y^k)^\top (x - y^k) - \frac{\mu}{2}\|x - y^k\|^2 - \frac{\alpha}{L}\nabla f(y^k)^\top \nabla f(z^k) + \frac{\alpha^2}{2L}\|\nabla f(z^k)\|^2 \\ &\quad - \frac{\tau}{L}\left(\nabla f(y^k)^\top \nabla f(z^k) - \|\nabla f(y^k)\|^2\right) - \frac{\eta\beta}{L}\|\nabla f(y^k)\|^2, \end{aligned} \tag{2.5.48}$$

for any $x \in \mathbb{R}^n$. At the same time, let us set $0 \le \eta < 1$, and we derive

$$\begin{aligned} \nabla f(y^k)^\top \nabla f(z^k) &= \nabla f(y^k)^\top \Big(\nabla f(y^k) + \nabla f(z^k) - \nabla f(y^k))\Big) \\ &\ge \|\nabla f(y^k)\|^2 - \|\nabla f(y^k)\| \cdot L\|z^k - y^k\| \\ &= (1-\eta)\|\nabla f(y^k)\|^2, \end{aligned} \tag{2.5.49}$$

and

$$\begin{aligned} \|\nabla f(z^k)\|^2 &\le \Big(\|\nabla f(y^k)\| + \|\nabla f(z^k) - \nabla f(y^k)\|\Big)^2 \\ &\le \Big(\|\nabla f(y^k)\| + L\|z^k - y^k\|\Big)^2 \\ &= (1+\eta)^2\|\nabla f(y^k)\|^2. \end{aligned} \tag{2.5.50}$$

Summing up (2.5.49) and (2.5.50), it follows from (2.5.48) that

$$\begin{aligned} f(x^{k+1}) \le\ & f(x) - \nabla f(y^k)^\top (x - y^k) - \frac{\mu}{2}\|x - y^k\|^2 \\ & - \frac{2\alpha(1-\eta) - (1+\eta)^2\alpha^2 + 2\eta(\beta - \tau)}{2L}\|\nabla f(y^k)\|^2. \end{aligned} \tag{2.5.51}$$

Taking $x = x^k$ and $x = x^*$ respectively, where x^* is the minimizer of f, we have

$$\begin{aligned} f(x^{k+1}) \le\ & f(x^k) - \nabla f(y^k)^\top (x^k - y^k) - \frac{\mu}{2}\|x^k - y^k\|^2 \\ & - \frac{2\alpha(1-\eta) - (1+\eta)^2\alpha^2 + 2\eta(\beta - \tau)}{2L}\|\nabla f(y^k)\|^2 \end{aligned} \tag{2.5.52}$$

and

$$\begin{aligned} f(x^{k+1}) \le\ & f(x^*) - \nabla f(y^k)^\top (x^* - y^k) - \frac{\mu}{2}\|x^* - y^k\|^2 \\ & - \frac{2\alpha(1-\eta) - (1+\eta)^2\alpha^2 + 2\eta(\beta - \tau)}{2L}\|\nabla f(y^k)\|^2. \end{aligned} \tag{2.5.53}$$

For a given $0 < \theta < 1$, let us multiply $1 - \theta$ on both sides of (2.5.52) and θ on both sides of (2.5.53), and then sum up the two inequalities. We obtain

$$\begin{aligned} & f(x^{k+1}) - f(x^*) \\ \le\ & (1-\theta)\Big(f(x^k) - f(x^*)\Big) + \nabla f(y^k)^\top \Big[(1-\theta)(y^k - x^k) + \theta(y^k - x^*)\Big] - \frac{\theta\mu}{2}\|x^* - y^k\|^2 \\ & - \frac{\mu(1-\theta)}{2}\|x^k - y^k\|^2 - \frac{2\alpha(1-\eta) - (1+\eta)^2\alpha^2 + 2\eta(\beta - \tau)}{2L}\|\nabla f(y^k)\|^2. \end{aligned} \tag{2.5.54}$$

Now, instead of trying to directly bound the last four terms on the right hand side of (2.5.54), the point v^k comes to help cancel them out, as shown below. Referring to (2.5.47),

let us choose $\omega_1 + \omega_2 = 1$. We have

$$\begin{aligned} & \|v^{k+1} - x^*\|^2 \\ \leq \; & \omega_1 \|v^k - x^*\|^2 + \omega_2 \|y^k - x^*\|^2 + \omega_3^2 \|\nabla f(y^k)\|^2 - 2\omega_3 \nabla f(y^k)^\top \left[\omega_1 v^k + \omega_2 y^k - x^*\right]. \end{aligned} \tag{2.5.55}$$

Selecting a parameter $C > 0$, and summing (2.5.54) with (2.5.55) (multiplying C), we have

$$\begin{aligned} & f(x^{k+1}) - f(x^*) + C\|v^{k+1} - x^*\|^2 \\ \leq \; & (1-\theta)(f(x^k) - f(x^*)) + C\omega_1 \|v^k - x^*\|^2 \\ & + \left(C\omega_2 - \frac{\theta\mu}{2}\right) \|y^k - x^*\|^2 \\ & + \left(C\omega_3^2 - \frac{2\alpha(1-\eta) - (1+\eta)^2\alpha^2 + 2\eta(\beta - \tau)}{2L}\right) \|\nabla f(y^k)\|^2 \\ & + \nabla f(y^k)^\top \left[(1-\theta)(y^k - x^k) + \theta(y^k - x^*) - 2\omega_3 C(\omega_1 v^k + \omega_2 y^k - x^*)\right]. \end{aligned} \tag{2.5.56}$$

Now we shall choose the parameters so that the last three terms in (2.5.56) are non-positive so that they can be dropped from the inequality. Summarizing, the requirements are:

$$\begin{cases} \theta & = & 2\omega_3 C \\ \nu_1 & = & \frac{1-\theta}{1-2\omega_2\omega_3 C} \\ \nu_2 & = & \frac{2\omega_1\omega_3 C}{1-2\omega_2\omega_3 C} \\ \eta & < & 1 \\ \omega_1 & \leq & 1-\theta \\ \omega_2 & = & 1-\omega_1 \\ \omega_2 C & \leq & \frac{\mu\theta}{2} \\ \omega_3^2 C & \leq & \frac{2\alpha(1-\eta)-(1+\eta)^2\alpha^2+2\eta(\beta-\tau)}{2L}. \end{cases} \tag{2.5.57}$$

In particular, for the last term in (2.5.56) we have:

$$\begin{aligned} & \nabla f(y^k)^\top \left[(1-\theta)(y^k - x^k) + \theta(y^k - x^*) - 2\omega_3 C(\omega_1 v^k + \omega_2 y^k - x^*)\right] \\ = \; & \nabla f(y^k)^\top \left[(1 - 2\omega_2\omega_3 C)y^k - (1-\theta)x^k - 2\omega_1\omega_3 C v^k\right] \\ = \; & \nabla f(y^k)^\top \left[(1 - 2\omega_2\omega_3 C)y^k - (1 - 2\omega_2\omega_3 C)p^k\right] \\ = \; & (1 - 2\omega_2\omega_3 C)\nabla f(y^k)^\top (y^k - p^k) \leq 0. \end{aligned}$$

We summarize our findings and arrive at the following theorem:

Theorem 2.5.1. *For an unconstrained optimization model* (2.5.44) *with the objective* $f(x)$ *being strongly convex with modulus* $\mu > 0$ *and gradient Lipschitz with constant* $L \geq \mu$*, the sequence* $\{x^k\}$*,* $k = 0, 1, 2, \ldots$ *generated by the extra-point scheme* (2.5.47) *converges linearly to the optimal solution:*

$$f(x^{k+1}) - f(x^*) + C\|v^{k+1} - x^*\|^2 \leq (1-\theta)\left(f(x^k) - f(x^*) + C\|v^k - x^*\|^2\right)$$

for properly chosen $0 < \theta < 1$ *and* C *and* ν_1, ν_2*,* $\alpha, \beta, \eta, \tau$*, and* $\omega_1, \omega_2, \omega_3$ *satisfying Condition* (2.5.57)*.*

The following parameter choices demonstrate an example to satisfy the condition (2.5.57), by taking $0 < \delta < 1$:

$$\left\{\begin{array}{l} \theta = \sqrt{\frac{\mu}{L}},\ \nu_1 = \frac{1}{1+\theta},\ \nu_2 = \frac{\theta}{1+\theta},\ \eta = \delta,\ \alpha = \frac{1-\delta}{(1+\delta)^2},\ \tau = \frac{1}{(1+\delta)^2}, \\ \beta = \frac{3}{(1+\delta)^2},\ \omega_1 = 1-\theta,\ \omega_2 = \theta,\ \omega_3 = \frac{1}{\sqrt{\mu L}},\ C = \frac{\mu}{2}. \end{array}\right. \tag{2.5.58}$$

If we further take $y^k = p^k$, procedure (2.5.47) can be simplified to

(2.5.59)

$$\left\{\begin{array}{lcl} y^k & := & \frac{x^k + \theta v^k}{1+\theta} \\ z^k & := & y^k - \frac{\delta}{L}\nabla f(y^k) \\ x^{k+1} & := & y^k - \frac{1-\delta}{(1+\delta)^2 L}\nabla f(z^k) - \frac{1}{(1+\delta)^2 L}\left(\nabla f(z^k) - \nabla f(y^k)\right) + \frac{3}{(1+\delta)^2}(z^k - y^k) \\ v^{k+1} & := & (1-\theta)v^k + \frac{\theta(\mu\delta - L)}{\mu\delta}y^k + \frac{\theta L}{\mu\delta}z^k, \end{array}\right.$$

and we have:

$$f(x^{k+1}) - f(x^*) + \frac{\mu}{2}\|v^{k+1} - x^*\|^2 \leq \left(1 - \sqrt{\sigma}\right)\left(f(x^k) - f(x^*) + \frac{\mu}{2}\|v^k - x^*\|^2\right).$$

The results are summarized in the next theorem:

Theorem 2.5.2. *Following up on Theorem 2.5.1, if we further specify the parameters as in* (2.5.58)*, then scheme* (2.5.47) *reduces to* (2.5.59)*. With the choice* $v^0 = x^0$*, the sequence* $\{x^k\}$ *converges linearly to* x^* *at the optimal rate*

$$f(x^k) - f(x^*) \leq 2\left(1 - \sqrt{\sigma}\right)^k \left(f(x^0) - f(x^*)\right).$$

Through an exemplifying scheme of the extra-point approach in optimization (2.5.47) and its more specific form (2.5.58) after fixing certain parameters, we show that it is also possible to construct a more general, but also optimal, first-order method under the given class of algorithms where the lower bound for (strongly) convex optimization is effective [66]. While the main structure of these schemes resemble that in Nesterov's accelerated method,

it is sufficient to fulfill the purpose of demonstrating the general extra-point concept widely used by various accelerated methods. Furthermore, the flexible forms of these schemes allow trainable parameter-tuning and structure learning in practical performance, which is the main motivation behind the proposed extra-point approach for designing optimal first-order methods.

2.6 Numerical Experiments

In the numerical experiments, we aim to test the performance of the three specialized extra-point schemes under different settings: strongly monotone VI, strongly monotone VI in combination with monotone gradient mapping, and strongly convex optimization. We especially compare the proposed schemes with first-order methods that can be viewed as specific configurations of the proposed schemes, in order to show the advantages of the flexibility in looking for configurations that lead to better numerical performance.

We first conduct two experiments under the VI setting, with the first one being unconstrained ($\mathcal{Z} = \mathbb{R}^n$) and the second being constrained by $z \geq 0$ ($\mathcal{Z} = \mathbb{R}^n_+$), both with a *linear* operator $F(z)$. Note that in the unconstrained case we are equivalently solving a linear equation system $F(z^*) = 0$, whereas in the constrained case the problem is equivalent to an LCP problem:

$$z^* \geq 0, \quad F(z^*) \geq 0, \quad (z^*)^\top F(z^*) = 0.$$

The linear strongly monotone operator $F(z)$ is set to be the same for both unconstrained and constrained cases, and is designed as the following:

$$F(z) = Mz + q, \tag{2.6.60}$$

where $M = Q + A$ is the sum of a positive diagonal matrix Q and a skew-symmetric matrix A, therefore a strongly monotone operator. The problem size in our experiment is $n = 20$, and the matrices Q, A, q are randomly generated. In particular, elements of Q are generated by three groups, uniformly distributed within $[0, 1]$, $[0, 10]$, and $[0, 50]$, respectively; A is generated by the difference of a uniformly distributed matrix and its transpose; q is generated by normal distribution $N(0, 10)$ elementwise.

We use gradient projection, extra-gradient method, and OGDA in comparison with the proposed extra-point scheme (2.3.14). We refer Table 2.1 for the parameters in the dynamics of each respective method. Figure 2.1 shows the convergence behavior for these methods. For the unconstrained case (the left plot in Figure 2.1), we use norm of the operator $\|F(z)\|$ as our measurement of convergence (merit function). On the other hand, we use $|z^\top F(z)|$ as

our measurement of convergence for the constrained case (the right plot in Figure 2.1). In both experiments, the proposed extra-point scheme has a superior convergence rate, followed by extra-gradient and OGDA methods, while in general the vanilla gradient projection method has the worst performance. All the methods are manually tuned to their best performance to a certain precision of the parameters.

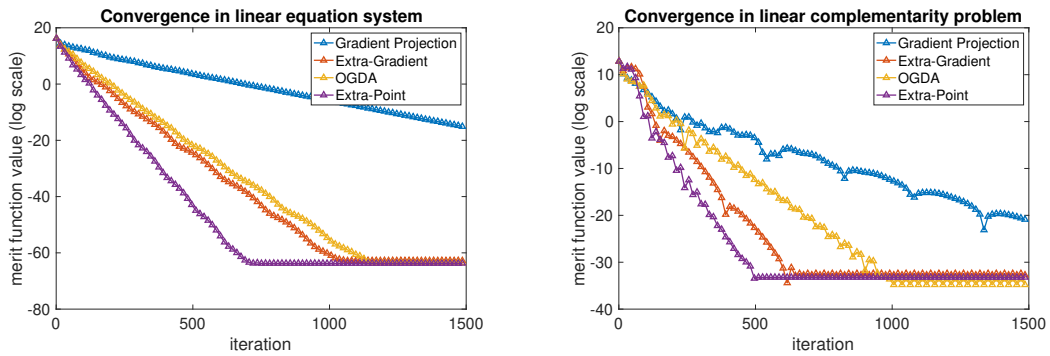

Figure 2.1: Convergence in strongly monotone VI

The next two experiments are conducted to demonstrate the convergence of the extra-point scheme (2.4.31) when a VI operator can be expressed as (2.4.29), which includes an explicit gradient mapping. The problem is set up as a constrained regularized saddle-point problem:

$$\begin{aligned}\min_{x\in\mathbb{R}^n_+}\max_{y\in\mathbb{R}^m_+} f(x,y) \quad &= \quad \frac{\lambda_1}{M_1}\sum_{i=1}^{M_1}\ln(1+e^{-a_i^\top x}) + \frac{1}{2}\|x\|^2 \\ &\quad +x^\top Ay - \frac{\lambda_2}{M_2}\sum_{j=1}^{M_2}\ln(1+e^{-b_j^\top y}) - \frac{1}{2}\|y\|^2.\end{aligned}$$

The corresponding VI operator can be expressed as:

$$F(z) = \begin{pmatrix} -\frac{\lambda_1}{M_1}\sum_{i=1}^{M_1}\frac{a_i}{1+e^{-a_i^\top x}} + x + Ay \\ -\frac{\lambda_2}{M_2}\sum_{j=1}^{M_2}\frac{b_j}{1+e^{-b_j^\top y}} + y - A^\top x \end{pmatrix}, \quad z := (x,y),$$

which admits to the following split of operator

$$F(z) = H(z) + \nabla g(z), \quad H(z) := \begin{pmatrix} I_n & A \\ -A^\top & I_m \end{pmatrix} z, \quad \nabla g(z) := \begin{pmatrix} -\frac{\lambda_1}{M_1}\sum_{i=1}^{M_1}\frac{a_i}{1+e^{-a_i^\top x}} \\ -\frac{\lambda_2}{M_2}\sum_{j=1}^{M_2}\frac{b_j}{1+e^{-b_j^\top y}} \end{pmatrix},$$

where $H(z)$ is strongly monotone with modulus 1 and $\nabla g(z)$ is monotone. In this formulation, we can exploit the constant λ_1, λ_2 to adjust the weights of the gradient mapping

$\nabla g(z)$. In particular, we set $\lambda_1 = \lambda_2 = \lambda$, $m = 2n = 100$, and $M_1 = M_2 = 100$. The elements in A, a_i, b_j are generated through standard normal distribution.

The convergence results are shown in Figure 2.2, where we compare the regular VI methods (gradient projection, extra-gradient, OGDA) with the proposed scheme (2.4.31) (denote by ExP-Acc) and the accelerated extra-gradient method (by setting $\beta = \gamma = \tau = 0$ in (2.4.31), which is the strongly monotone variant of the accelerated mirror-prox method in [10], denote by EG-Acc). We use $\frac{1}{2}\|z - P_{\mathcal{Z}}(z - \frac{1}{\mu}F(z))\|^2$ as the merit function, where $\mathcal{Z} := \mathbb{R}^n_+ \times \mathbb{R}^m_+$. As shown in the plots (left: $\lambda = 50$; right: $\lambda = 200$), regular accelerated VI methods (extra-gradient, OGDA) could perform worse than the gradient projection method, while the accelerated methods making explicit use of the gradient mapping (ExP-Acc, EG-Acc) outperform others. Moreover, being more flexible in finding various configurations as featured in the extra-point approach, ExP-Acc shows further improvement over EG-Acc.

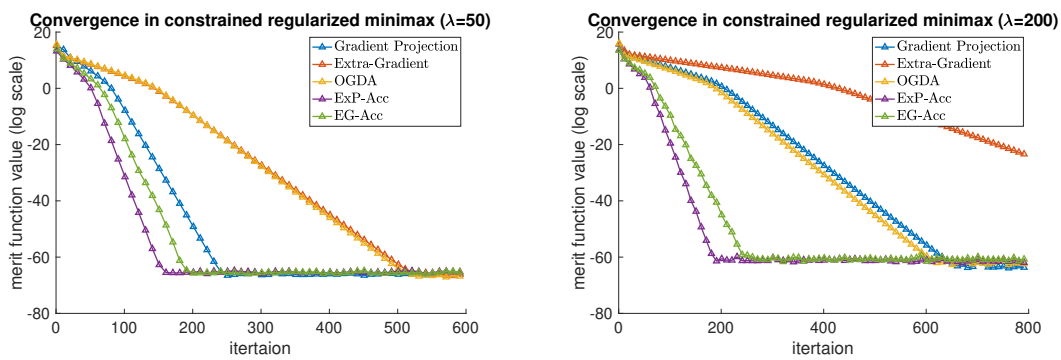

Figure 2.2: Convergence in strongly monotone VI combined with gradient mapping

We also conduct two experiments under the optimization setting, with the following unconstrained *regularized logistic regression* model:

$$\min_x \frac{1}{M}\sum_{i=1}^{M} \ln(1 + e^{-a_i^\top x}) + \frac{\rho}{2}\|x\|^2.$$

Note that although the objection function is strongly convex with modulus $\mu = \rho$, the Lipschitz constant needs to be estimated for different problem parameters. We set $\rho = 0.005$ in both experiments and generate the elements in a_i with normal distribution $N(0, 5)$. The number of data points M and the problem size n are varied for the two experiments.

In both experiments, we compare the specialized extra-point scheme for optimization (2.5.47) (setting $y^k = p^k$) with other first-order methods. Note that while extra-gradient method and OGDA are not guaranteed to be optimal in optimization, we include their results mainly for the purpose of observing their convergence in optimization. The convergence results for each method are shown in Figure 2.3, and the progress is measured as the

norm of gradient. In these experiments, the optimal methods for VI (extra-gradient method and OGDA) could perform worse than the vanilla gradient descent. Nesterov's accelerated method and the proposed extra-point scheme have the best performance among these first-order methods. In addition, the extra-point scheme shows a even faster convergence over Nesterov's method after fine-tuning the parameters.

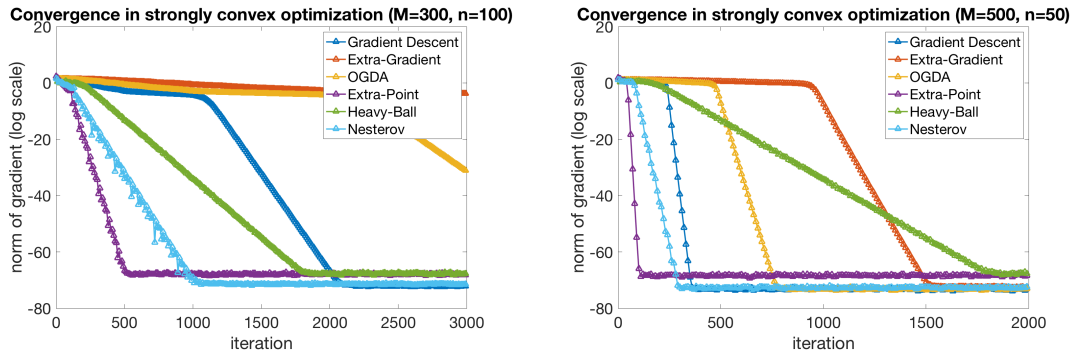

Figure 2.3: Convergence in strongly convex optimization

2.7 Conclusion

In this chapter, we propose a general extra-point approach capable of guiding the design of specific first-order methods within the scope under which the lower-bound results take effect [66, 108]. By building on the idea of extra points that has been observed in accelerated/optimal first-order methods, the extra-point approach provides a structured and flexible framework to develop algorithms whose iteration complexity matches the lower bounds, hence optimal. We present three applications of the extra-point approach through specialized schemes for strongly monotone VI, strongly monotone VI in combination with monotone gradient mapping, and strongly convex optimization, respectively. In these schemes, we use the ideas and search directions that have been proposed in the literature in order to demonstrate the constructions of the extra-point sequences. The newly proposed schemes turn out to be more general and flexible than the existing known accelerated methods, which use only the individual components of the general schemes. We then show that the iteration complexities of these schemes are guaranteed to be optimal even for much more flexible configurations, and such flexibility is further shown to be useful to enhance the numerical performance.

Chapter 3

An Approximation-Based Regularized Extra-Gradient Method for Monotone Variational Inequality Problems

In this chapter, we propose a general extra-gradient scheme for solving monotone variational inequalities (VI), referred to here as *Approximation-based Regularized Extra-gradient method* (ARE). The first step of ARE solves a VI subproblem with an approximation operator satisfying a pth-order Lipschitz bound with respect to the original mapping, further coupled with the gradient of a $(p+1)^{\text{th}}$-order regularization. The optimal global convergence is guaranteed by including an additional extra-gradient step, while a pth-order superlinear local convergence is shown to hold if the VI is strongly monotone. The proposed ARE is inclusive and general, in the sense that a variety of solution methods can be formulated within this framework as different manifestations of approximations, and their iteration complexities would follow through in a unified fashion. The ARE framework relates to the first-order methods, while opening up possibilities to developing higher-order methods specifically for structured problems that guarantee the optimal iteration complexity bounds.

3.1 Preliminaries

For the VI problem (1.1.1), as a notation we denote the solution set as

$$\mathrm{VI}_{\mathcal{X}}(F(x)) := \{x^* \mid x^* \in \mathcal{X} \text{ such that } F(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in \mathcal{X}\}.$$

Let us first introduce a few terminologies that will be used throughout the chapter. The term "p^{th}-order method" will be used following the convention of optimization. In particular, by considering $F(x) = \nabla f(x)$ specifically as a gradient mapping of some function $f(x)$, the first-order method in VI refers to using only the information from the operator $F(\cdot)$, and the pth-order method refers to using the $(p-1)^{th}$-order derivative of the operator: $\nabla^{p-1}F$. As a result, the term "gradient" will also be used to refer to $F(\cdot)$ due to the background of VI in solving saddle-point and optimization models. In the pth-order method, the Lipschitz continuity of $\nabla^{p-1}F(x)$ is assumed with constant L_p:

$$\|\nabla^{p-1}F(x) - \nabla^{p-1}F(y)\| \leq L_p\|x - y\|. \tag{3.1.1}$$

In this chapter, the proposed *Approximation-based Regularized Extra-gradient method* (ARE) can be viewed as a generalization of the extra-gradient method (1.1.6) in some sense. While the intermediate iterate $x^{k+0.5}$ in extra-gradient method (1.1.6) is updated by a gradient projection step, in ARE it is replaced by solving a VI subproblem:

$$x^{k+0.5} := \text{VI}_{\mathcal{X}}(\tilde{F}(x; x^k) + \gamma\|x - x^k\|^{p-1}(x - x^k)), \tag{3.1.2}$$

where $\tilde{F}(x; x^k)$ is an approximation mapping at x^k that satisfies a pth-order Lipschitz bound with respect to $F(x)$ (will be formally defined later), and $\|x-x^k\|^{p-1}(x-x^k)$ is the gradient mapping of a $(p+1)^{\text{th}}$-order regularization. Therefore, we refer to the update in (3.1.2) as $(p+1)^{th}$*-order regularized VI subproblem*. A common choice of $\tilde{F}(x; x^k)$ is the *Taylor approximation* of $F(x)$ at x^k, namely,

$$\tilde{F}(x; x^k) := \sum_{i=0}^{p-1} \frac{1}{i!}\nabla^i F(x^k)[x - x^k]^i.$$

Such choice of $\tilde{F}(x; x^k)$ not only recovers the extra-gradient method when $p = 1$, but also gives a succinct update principle for higher-order methods when $p > 1$. However, the Taylor approximation needs not be the only motivation for ARE. We show that the key underlying condition is the aforementioned p^{th}-order Lipschitz bound, therefore any approximation satisfying this condition can be considered as a valid method under the general framework of ARE. This not only generalizes the existing methods but also opens up the possibilities of developing different methods from those in the literature, and we will discuss several such specific schemes in Section 3.5. By applying the abstraction of "approximation" in ARE, a unifying and concise analysis is available to establish the iteration complexity bound that can be readily specified to any concrete approximation in various methods given the different problem structures at hand.

3.2 The Global Convergence Analysis of ARE

The Approximation-based Regularized Extra-gradient method (ARE) aims to solve the VI problem:

$$\text{(3.2.3)}\ \mathrm{VI}_{\mathcal{X}}(F(x)) := \{x^* \mid x^* \in \mathcal{X} \text{ such that } F(x^*)^\top (x - x^*) \geq 0 \text{ for all } x \in \mathcal{X}\}.$$

We assume that $F(x)$ is monotone (1.1.2) and $\mathrm{VI}_{\mathcal{X}}(F(x))$ is non-empty. When $F(x)$ is assumed to be strongly monotone, $\mathrm{VI}_{\mathcal{X}}(F(x))$ becomes a singleton. We also assume the pth-order Lipschitz continuity (3.1.1).

Now, given an arbitrary $y \in \mathcal{X}$, we are interested in a general approximation mapping at y: $\tilde{F}(\cdot; y) : \mathbb{R}^n \mapsto \mathbb{R}^n$, such that the following p^{th}-order Lipschitz bound holds between the original mapping $F(x)$ and the approximation $\tilde{F}(x; y)$:

$$\|\tilde{F}(x; y) - F(x)\| \leq \tau L_p \|x - y\|^p, \tag{3.2.4}$$

for some $p > 1$ and $\tau \in (0, 1]$. The examples of such approximation include but not limited to the general Taylor approximation, which further includes $\tilde{F}(x; y) = F(y)$ for $p = 1$ and $\tilde{F}(x; y) = F(y) + \nabla F(y)(x - y)$ for $p = 2$ as special cases. In general, we say the proposed ARE is a "p^{th}-order method" if the Lipschitz bound (3.2.4) holds with p.

Based on the approximation mapping $\tilde{F}(x; y)$, let us consider the *regularized* approximation mapping by adding a gradient mapping of the $(p+1)^{\text{th}}$-order regularization term, expressed in the following form:

$$\tilde{F}(x; y) + L_p \|x - y\|^{p-1}(x - y). \tag{3.2.5}$$

Since the Jacobian of $L_p\|x - y\|^{p-1}(x - y)$ is positive definite for $x \neq y$, the mapping is monotone [19]. Therefore, the regularized approximation mapping (3.2.5) is also monotone as long as $\tilde{F}(x; y)$ is. In ARE, the first step in each iteration is solving a VI subproblem with operator (3.2.5), i.e. a $(p + 1)^{\text{th}}$-order regualarized VI subproblem, followed by an extra-gradient step, summarized as follows:

$$\text{(3.2.6)} \begin{cases} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(\tilde{F}(x; x^k) + L_p\|x - x^k\|^{p-1}(x - x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x \in \mathcal{X}} \langle F(x^{k+0.5}), x - x^k\rangle + \frac{L_p\|x^{k+0.5} - x^k\|^{p-1}}{2}\|x - x^k\|^2, \end{cases}$$

for $k = 1, 2, \ldots$. In the above, as a matter of notion we indicate $x^{k+0.5}$ to be any solution taken from the solution set $\mathrm{VI}_{\mathcal{X}}(\cdot)$. In the second step, the extra-gradient step involves a varying step size $\frac{1}{\gamma_k}$, where

$$\gamma_k := L_p\|x^{k+0.5} - x^k\|^{p-1}, \quad k \geq 1$$

is a parameter depending on the previous update $x^{k+0.5}$. This together with the bound (3.2.4) form the basis of the optimal iteration complexity bound for ARE.

3.2.1 Solving monotone VI problems with ARE

We first establish the global convergence results for solving a general monotone VI (3.2.3) with p^{th}-order ARE in the next theorem.

Theorem 3.2.1 (Global convergence of ARE: Monotone VI)**.** *Let $\{x^k\}_{k\geq 1}$ and $\{x^{k+0.5}\}_{k\geq 1}$ be generated by (3.2.6) and suppose $F(\cdot)$ is monotone and $\tilde{F}(x;x^k)$ is such that (3.2.4) holds. Then*

$$m(\bar{x}_N) := \max_{x\in\mathcal{X}} \langle F(x), \bar{x}_N - x\rangle \leq \frac{D^2}{2\Gamma_N} = \mathcal{O}(N^{-\frac{p+1}{2}}), \tag{3.2.7}$$

where

$$\bar{x}_N := \frac{\sum_{k=1}^{N} \frac{x^{k+0.5}}{\gamma_k}}{\Gamma_N}, \quad \Gamma_N := \sum_{k=1}^{N} \gamma_k^{-1}$$

for some $N > 0$, and $D := \max_{x,x'\in\mathcal{X}} \|x - x'\|$.

Proof. Since

$$x^{k+0.5} = \mathrm{VI}_{\mathcal{X}}\left(\tilde{F}(x;x^k) + L_p\|x - x^k\|^{p-1}(x - x^k)\right),$$

we have

$$\langle \tilde{F}(x^{k+0.5};x^k) + L_p\|x^{k+0.5} - x^k\|^{p-1}(x^{k+0.5} - x^k), x - x^{k+0.5}\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{3.2.8}$$

Denote $\gamma_k = L_p\|x^{k+0.5} - x^k\|^{p-1}$. Substituting $x = x^{k+1}$ in (3.2.8) we have

$$\begin{aligned} & \langle \tilde{F}(x^{k+0.5};x^k), x^{k+1} - x^{k+0.5}\rangle \\ \geq\ & \gamma_k\langle x^{k+0.5} - x^k, x^{k+0.5} - x^{k+1}\rangle \\ =\ & \frac{\gamma_k}{2}\left(\|x^{k+0.5} - x^k\|^2 + \|x^{k+1} - x^{k+0.5}\|^2 - \|x^{k+1} - x^k\|^2\right). \end{aligned} \tag{3.2.9}$$

On the other hand, by the optimality condition at x^{k+1} we have

$$\langle F(x^{k+0.5}) + \gamma_k(x^{k+1} - x^k), x - x^{k+1}\rangle \geq 0, \text{ for all } x \in \mathcal{X}.$$

Hence,

$$\begin{aligned} & \langle F(x^{k+0.5}), x - x^{k+1}\rangle \\ \geq\ & \gamma_k\langle x^{k+1} - x^k, x^{k+1} - x\rangle \\ =\ & \frac{\gamma_k}{2}\left(\|x^{k+1} - x\|^2 + \|x^{k+1} - x^k\|^2 - \|x^k - x\|^2\right), \text{ for all } x \in \mathcal{X}. \end{aligned} \tag{3.2.10}$$

Continue with the above inequality, for any given $x \in \mathcal{X}$ we have

$$
\begin{aligned}
& \frac{\gamma_k}{2}\left(\|x^{k+1}-x\|^2+\|x^{k+1}-x^k\|^2-\|x^k-x\|^2\right) \\
\overset{(3.2.10)}{\leq} \quad & \langle F(x^{k+0.5}), x-x^{k+1}\rangle \\
= \quad & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \langle F(x^{k+0.5}), x^{k+0.5}-x^{k+1}\rangle \\
= \quad & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \langle F(x^{k+0.5})-\tilde{F}(x^{k+0.5};x^k), x^{k+0.5}-x^{k+1}\rangle \\
& +\langle \tilde{F}(x^{k+0.5};x^k), x^{k+0.5}-x^{k+1}\rangle \\
\leq \quad & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \|F(x^{k+0.5})-\tilde{F}(x^{k+0.5};x^k)\|\cdot\|x^{k+0.5}-x^{k+1}\| \\
& +\langle \tilde{F}(x^{k+0.5};x^k), x^{k+0.5}-x^{k+1}\rangle \\
\leq \quad & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \frac{\|F(x^{k+0.5})-\tilde{F}(x^{k+0.5};x^k)\|^2}{2\gamma_k} + \frac{\gamma_k\|x^{k+0.5}-x^{k+1}\|^2}{2} \\
& +\langle \tilde{F}(x^{k+0.5};x^k), x^{k+0.5}-x^{k+1}\rangle \\
\overset{(3.2.4)}{\leq} \quad & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \frac{\tau^2 L_p^2\|x^{k+0.5}-x^k\|^{2p}}{2\gamma_k} + \frac{\gamma_k\|x^{k+0.5}-x^{k+1}\|^2}{2} \\
& +\langle \tilde{F}(x^{k+0.5};x^k), x^{k+0.5}-x^{k+1}\rangle.
\end{aligned}
$$

Noticing that $\frac{\tau^2 L_p^2\|x^{k+0.5}-x^k\|^{2p}}{2\gamma_k} = \frac{\tau^2\gamma_k\|x^{k+0.5}-x^k\|^2}{2}$, and further using (3.2.9) we derive from the above that

$$
\begin{aligned}
& \frac{\gamma_k}{2}\left(\|x^{k+1}-x\|^2+\|x^{k+1}-x^k\|^2-\|x^k-x\|^2\right) \\
\leq \quad & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \frac{\tau^2\gamma_k\|x^{k+0.5}-x^k\|^2}{2} + \frac{\gamma_k\|x^{k+0.5}-x^{k+1}\|^2}{2} \\
& +\frac{\gamma_k}{2}\left[-\|x^{k+0.5}-x^k\|^2-\|x^{k+1}-x^{k+0.5}\|^2+\|x^{k+1}-x^k\|^2\right].
\end{aligned}
$$

Canceling out terms, we simplify the above inequality into

(3.2.11)
$$
\langle F(x^{k+0.5}), x^{k+0.5}-x\rangle + \frac{\gamma_k}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \leq \frac{\gamma_k}{2}\left[\|x^k-x\|^2-\|x^{k+1}-x\|^2\right].
$$

Consequently, by the monotonicity of F, we have

$$
\begin{aligned}
& \langle F(x), x^{k+0.5}-x\rangle + \frac{\gamma_k}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \\
\leq \quad & \langle F(x^{k+0.5}), x^{k+0.5}-x\rangle + \frac{\gamma_k}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \\
\leq \quad & \frac{\gamma_k}{2}\left[\|x^k-x\|^2-\|x^{k+1}-x\|^2\right].
\end{aligned}
$$

Dividing both sides by γ_k yields

(3.2.12)
$$
\frac{1}{\gamma_k}\left\langle F(x), x^{k+0.5}-x\right\rangle + \frac{1}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \leq \frac{1}{2}\left[\|x^k-x\|^2-\|x^{k+1}-x\|^2\right].
$$

Summing up the inequality (3.2.12) from $k = 1$ to N, and dividing the resulting inequality on both sides by Γ_N we obtain

$$\langle F(x), \bar{x}_N - x\rangle + \frac{1-\tau^2}{2\Gamma_N}\sum_{k=1}^{N}\|x^{k+0.5} - x^k\|^2 \leq \frac{\|x^1 - x\|^2}{2\Gamma_N} \tag{3.2.13}$$

for any $x \in \mathcal{X}$. Taking $x = x^*$ in (3.2.13) yields

$$\sum_{k=1}^{N}\|x^{k+0.5} - x^k\|^2 \leq \frac{\|x^1 - x^*\|^2}{1-\tau^2}. \tag{3.2.14}$$

The so-called *mean inequality* maintains that for any positive sequence $\{a_k > 0 \mid k = 1, 2, ..., N\}$ and any real value r, if we define

$$M_r(a) := \left(\frac{1}{N}\sum_{k=1}^{N} a_k^r\right)^{\frac{1}{r}}$$

then we have $M_{r_1}(a) \leq M_{r_2}(a)$ for any $r_1 \leq r_2$.

Now, if we let $a_k := \|x^{k+0.5} - x^k\|^{-(p-1)}$, then we have $M_{-\frac{2}{p-1}}(a) \leq M_1(a)$; that is

$$\left(\frac{1}{N}\sum_{k=1}^{N}\|x^{k+0.5} - x^k\|^2\right)^{-\frac{p-1}{2}} \leq \frac{1}{N}\sum_{k=1}^{N}\|x^{k+0.5} - x^k\|^{-(p-1)}.$$

Therefore,

$$\begin{aligned}\Gamma_N &= \frac{1}{L_p}\sum_{k=1}^{N}\|x^{k+0.5} - x^k\|^{-(p-1)} \geq \frac{N}{L_p}\left(\frac{1}{N}\sum_{k=1}^{N}\|x^{k+0.5} - x^k\|^2\right)^{-\frac{p-1}{2}} \\ &\overset{(3.2.14)}{\geq} \frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{\|x^1 - x^*\|^2}\right)^{\frac{p-1}{2}} \geq \frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{D^2}\right)^{\frac{p-1}{2}}.\end{aligned}$$

Then, (3.2.13) leads to

$$m(\bar{x}_N) \leq \frac{D^2}{2\Gamma_N} \leq \frac{D^2 L_p}{2N^{\frac{p+1}{2}}\left(\frac{1-\tau^2}{D^2}\right)^{\frac{p-1}{2}}} = \frac{D^{p+1}L_p}{2\left(1-\tau^2\right)^{\frac{p-1}{2}} N^{\frac{p+1}{2}}} = O\left(\frac{1}{N^{\frac{p+1}{2}}}\right).$$

□

Theorem 3.2.1 implies that the proposed ARE generated $\bar{x}_N$ such that $m(\bar{x}_N) \leq \epsilon$ with iteration complexity $\mathcal{O}\left(1/\epsilon^{\frac{2}{p+1}}\right)$. This matches the lower bound $\Omega\left(1/\epsilon^{\frac{2}{p+1}}\right)$ established in [53], hence optimal. The concept of solving a $(p+1)^{\text{th}}$-order regularized VI subproblem is

also proposed in [1] and [53], therefore there is no need for an additional bisection subroutine in each iteration. The major difference between ARE and the method proposed in [1] is that ARE uses a more general approximation operator $\tilde{F}(x; x^k)$ in the aforementioned regularized VI subproblem, which generalizes the Taylor approximation proposed in [1], and we provide a unified analysis as long as (3.2.4) is satisfied. The flexibility of not being restricted to Taylor approximation is demonstrated in Section 3.4.2 and Section 3.5, where examples are given for applying ARE with non-Taylor approximation $\tilde{F}(x; x^k)$ to problems when the original operator exhibits composite structure $F(x) = H(x) + G(x)$ or more generally $F(x) = H(G(x))$.

3.2.2 Solving strongly monotone VI problems with ARE-restart

While Theorem 3.2.1 establishes the optimal sublinear convergence for monotone VI, we shall incorporate a *restart* procedure into ARE (3.2.6) to further establish an improved linear convergence for strongly monotone VI. Similar restarting procedure is also seen in previous work [73, 53] for establishing the linear convergence. Below we give a detailed analysis for restarting the pth-order ARE (3.2.6), referred to as *ARE-restart*.

The ARE-restart works in *epochs*. That is, for each epoch m, where $m = 1, 2, ...$, a number of iterative updates (3.2.6) is performed and the output is set as the initial iterate at the start of the next epoch.

Let N_m denote the number of iterations performed in m^{th} epoch, $m = 1, 2, ...$ After each N_m iterations (of ARE), we restart ($x^1 \leftarrow \bar{x}_{N_m}$) and proceed to $(m+1)^{\text{th}}$ epoch. The output of epoch m is defined as:

$$\bar{x}_{N_m} := \frac{\sum_{k=1}^{N_m} \frac{1}{\gamma_k} x^{k+0.5}}{\Gamma_{N_m}} \in \mathcal{X}, \text{ and } \Gamma_{N_m} := \sum_{k=1}^{N_m} \frac{1}{\gamma_k}.$$

Denote $D_0 = \|x^1 - x^*\|$ as the distance to the solution from the very first initial iterate before any restart and note that $D_0 \leq D$. Let $0 < \delta < 1$ be a constant independent of the problem. Let us fix the iterations in each epoch:

$$N_1 = N_2 = \cdots = N = \left(\frac{L_p}{2\delta\mu}\right)^{\frac{2}{p+1}} \left(\frac{D_0^2}{1-\tau^2}\right)^{\frac{p-1}{p+1}}. \tag{3.2.15}$$

From the analysis in Theorem 3.2.1, we can first reach (3.2.11), where by using the strong

monotonicity we have:

$$
\begin{aligned}
& \mu\|x^{k+0.5}-x\|^2+\langle F(x),x^{k+0.5}-x\rangle+\frac{\gamma_k}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \\
\leq\ & \langle F(x^{k+0.5}),x^{k+0.5}-x\rangle+\frac{\gamma_k}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \\
\leq\ & \frac{\gamma_k}{2}\left[\|x^k-x\|^2-\|x^{k+1}-x\|^2\right]. \qquad (3.2.16)
\end{aligned}
$$

Taking $x=x^*$ and sum the inequality from $k=1$ to N:

$$
\begin{aligned}
& \mu\Gamma_N\|\bar{x}_N-x^*\|^2 \\
\leq\ & \sum_{k=1}^{N}\frac{\mu}{\gamma_k}\|x^{k+0.5}-x^*\|^2 \\
\leq\ & \frac{1}{2}\|x^1-x^*\|^2-\frac{1}{2}\|x^{N+1}-x^*\|^2-\frac{1-\tau^2}{2}\sum_{k=1}^{N}\|x^{k+0.5}-x^k\|^2, \qquad (3.2.17)
\end{aligned}
$$

where the first inequality is due to the convexity of the squared norm function.

Now consider the first epoch, inequality (3.2.17) implies:

$$\|\bar{x}_{N_1}-x^*\|^2\leq\frac{1}{2\mu\Gamma_{N_1}}D_0^2,$$

where

$$\Gamma_{N_1}\geq\frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{\|x^1-x^*\|^2}\right)^{\frac{p-1}{2}}=\frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{D_0^2}\right)^{\frac{p-1}{2}}.$$

Therefore,

$$\|\bar{x}_{N_1}-x^*\|^2\leq\frac{1}{2\mu\Gamma_{N_1}}D_0^2\leq\frac{L_p}{2\mu}\frac{D_0^{p+1}}{\left(1-\tau^2\right)^{\frac{p-1}{2}}}\frac{1}{N^{\frac{p+1}{2}}}=\delta\cdot D_0^2.$$

Now, in the second epoch, we take $x^1\leftarrow\bar{x}_{N_1}$. Similarly, we have:

$$\|\bar{x}_{N_2}-x^*\|^2\leq\frac{1}{2\mu\Gamma_{N_2}}\|\bar{x}_{N_1}-x^*\|^2\leq\frac{\delta}{2\mu\Gamma_{N_2}}D_0^2,$$

where the lower bound of Γ_{N_2} can also be estimated from (3.2.17), with an improved distance to solution $\|x^1-x^*\|^2=\|\bar{x}_{N_1}-x^*\|^2\leq\delta D_0^2$:

$$\Gamma_{N_2}\geq\frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{\|\bar{x}_{N_1}-x^*\|^2}\right)^{\frac{p-1}{2}}\geq\frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{\delta D_0^2}\right)^{\frac{p-1}{2}}.$$

Note that in the second epoch, the lower bound of Γ_{N_2} is improved by $\left(\frac{1}{\delta}\right)^{\frac{p-1}{2}}$. Then we have:

$$\|\bar{x}_{N_2}-x^*\|^2\leq\frac{\delta}{2\mu\Gamma_{N_2}}D_0^2\leq\frac{L_p}{2\mu}\frac{D_0^{p+1}}{\left(1-\tau^2\right)^{\frac{p-1}{2}}}\frac{\delta^{\frac{p+1}{2}}}{N^{\frac{p+1}{2}}}=\delta^{\frac{p+3}{2}}\cdot D_0^2.$$

Note that after second epoch, the distance is not decreased by a factor of δ^2 but a factor of $\delta^{\frac{p+3}{2}}$ instead. This is because while performing N iterations in one epoch provides a decrease of δ in terms of the original distance D_0, starting from an iterate $x^1 = \bar{x}^{N_1}$ in the second epoch provides an additional decrease of $\delta^{\frac{p+1}{2}}$ due to a better bound for $\|\bar{x}_{N_1} - x^*\|^2$ and Γ_{N_2}. Now continue considering the third epoch:

$$\|\bar{x}_{N_3} - x^*\|^2 \leq \frac{1}{2\mu\Gamma_{N_3}}\|\bar{x}_{N_2} - x^*\|^2 \leq \frac{\delta^{\frac{p+3}{2}}}{2\mu\Gamma_{N_3}} D_0^2,$$

where

$$\Gamma_{N_3} \geq \frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{\|\bar{x}_{N_2} - x^*\|^2}\right)^{\frac{p-1}{2}} \geq \frac{N^{\frac{p+1}{2}}}{L_p}\left(\frac{1-\tau^2}{\delta^{\frac{p+3}{2}} D_0^2}\right)^{\frac{p-1}{2}}.$$

Therefore,

$$\|\bar{x}_{N_3} - x^*\|^2 \leq \frac{\delta^{\frac{p+3}{2}}}{2\mu\Gamma_{N_3}} D_0^2 \leq \frac{L_p}{2\mu}\frac{D_0^{p+1}}{(1-\tau^2)^{\frac{p-1}{2}}}\frac{\delta^{\frac{(p+3)(p+1)}{4}}}{N^{\frac{p+1}{2}}} = \delta^{\frac{(p+3)(p+1)}{4}+1} \cdot D_0^2.$$

To summarize, after m epochs, we have

$$\|\bar{x}_{N_m} - x^*\|^2 \leq \delta^{t_m} \cdot D_0^2,$$

where

$$t_m = t_{m-1} \cdot \frac{p+1}{2} + 1, \quad t_1 = 1.$$

Then we have

$$\|\bar{x}_{N_m} - x^*\|^2 \leq \delta^{t_m} \cdot D_0^2 \leq \delta^{\left(\frac{p+1}{2}\right)^{m-1}} \cdot D_0^2 \leq \delta^{\left(\frac{p+1}{2}\right)^{m-1}} \cdot D^2.$$

That is, the total number of epochs required to have $\|\bar{x}_{N_m} - x^*\|^2 \leq \epsilon$ is given by

$$\log_{\frac{p+1}{2}} \log_{\frac{1}{\delta}} \frac{D^2}{\epsilon}, \tag{3.2.18}$$

for $p > 1$, a superlinear rate for the *epochs*. For $p = 1$, the log is one layer and we only have linear convergence. Note that, however, (3.2.18) is only the number of *epochs* needs to be run, and for each epoch a fixed number of N iterations is still performed, so the total iteration complexity is:

$$\left(\frac{L_p}{2\delta\mu}\right)^{\frac{2}{p+1}}\left(\frac{D_0^2}{1-\tau^2}\right)^{\frac{p-1}{p+1}} \log_{\frac{p+1}{2}} \log_{\frac{1}{\delta}} \frac{D^2}{\epsilon}$$

for $p > 1$. For simplicity, we can take $\delta = \frac{1}{2}$ and replace D_0 with D in the number of iterations N in one epoch, which gives the complexity:

$$\mathcal{O}\left(\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}} (D^2)^{\frac{p-1}{p+1}} \log_{\frac{p+1}{2}} \log_2 \frac{D^2}{\epsilon}\right).$$

The result is summarized in the next theorem.

Theorem 3.2.2 (Global convergence of ARE: Strongly monotone VI)**.** *Let* $\{x^k\}_{k\geq 1}$ *and* $\{x^{k+0.5}\}_{k\geq 1}$ *be generated by ARE* (3.2.6) *and suppose* $F(\cdot)$ *is strongly monotone with* $\mu > 0$ *and* $\tilde{F}(x; x^k)$ *is such that* (3.2.4) *holds. By restarting ARE after each* $N_i = N$ *iterations in epoch* i*, where* N *is given by* (3.2.15)*, the total number of epochs* m *required to obtain an output* $\bar{x}_{N_m}$ *such that* $\|\bar{x}_{N_m} - x^*\|^2 \leq \epsilon$ *is given by:*

$$\begin{cases} \mathcal{O}\left(\log_{\frac{p+1}{2}} \log_2 \frac{D^2}{\epsilon}\right), & p > 1, \\ \mathcal{O}\left(\log_2 \frac{D^2}{\epsilon}\right), & p = 1. \end{cases}$$

The total iteration complexity mN *is given by:*

$$\begin{cases} \mathcal{O}\left(\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}} (D^2)^{\frac{p-1}{p+1}} \log_{\frac{p+1}{2}} \log_2 \frac{D^2}{\epsilon}\right), & p > 1, \\ \mathcal{O}\left(\frac{L_1}{\mu} \log_2 \frac{D^2}{\epsilon}\right), & p = 1. \end{cases} \tag{3.2.19}$$

Through a careful analysis of the restarting procedure, Theorem 3.2.2 shows that by restarting ARE when $F(x)$ is strongly monotone, the optimal iteration complexity is achievable for $p = 1$ and the improved iteration complexity is obtained for $p > 1$, as summarized in (3.2.19). We note that the total number of epochs (or the number of restarting) is only of the order $\mathcal{O}\left(\log_{\frac{p+1}{2}} \log_2 \frac{D^2}{\epsilon}\right)$, which is an improved bound compared to $\mathcal{O}\left(\log_2 \frac{D^2}{\epsilon}\right)$ established in [73, 53]. This implies that the output iterate $\bar{x}_{N_j}$ after each epoch for $j = 1, ..., m$ converges towards x^* at a superlinear rate, and the reason being that the lower bound for the averaging parameter Γ_{N_j} is improved by an order of $\frac{p+1}{2}$.

3.3 The Local Convergence Analysis of ARE

In this section we shall analyze the local convergence behavior of ARE for strongly monotone F (i.e. $\mu > 0$) when $p > 1$. A pure Newton method typically exhibits local quadratic convergence in optimization, and the same has been shown for VI [92, 25]. In [73], while the global iterations proceed with restarting higher-order mirror-prox method [8] and an iteration complexity similar to (3.2.19) is established, the local iterations are performed by adopting CRN-SPP [25] to obtain quadratic convergence. The local superlinear convergence is further improved in [53] by restarting Perseus and in [33], to the order $\frac{p+1}{2}$. In the following analysis, we show that in the pth-order ARE where (3.2.4) is satisfied, then the pth-order local superlinear convergence rate holds, which is an improvement compared to existing work in the literature.

We first show that for the ARE update (3.2.6), $\|x^{k+0.5} - x^*\|$ converges to zero p^{th} -order superlinearly compared to $\|x^k - x^*\|$, as long as $\|x^{k+0.5} - x^k\|$ is sufficiently small. By the definition of a VI solution for update $x^{k+0.5}$:

$$\langle \tilde{F}(x^{k+0.5}; x^k) + L_p\|x^{k+0.5} - x^k\|^{p-1}(x^{k+0.5} - x^k), x - x^{k+0.5}\rangle \geq 0, \ \forall x \in \mathcal{X}.$$

Then

$$\begin{aligned}
& \langle F(x^{k+0.5}), x^{k+0.5} - x\rangle \\
\leq \quad & \langle \tilde{F}(x^{k+0.5}; x^k) - F(x^{k+0.5}) + L_p\|x^{k+0.5} - x^k\|^{p-1}(x^{k+0.5} - x^k), x - x^{k+0.5}\rangle \\
\leq \quad & \left\|\tilde{F}(x^{k+0.5}; x^k) - F(x^{k+0.5}) + L_p\|x^{k+0.5} - x^k\|^{p-1}(x^{k+0.5} - x^k)\right\| \cdot \|x - x^{k+0.5}\| \\
\leq \quad & \left(\left\|\tilde{F}(x^{k+0.5}; x^k) - F(x^{k+0.5})\right\| + \left\|L_p\|x^{k+0.5} - x^k\|^{p-1}(x^{k+0.5} - x^k)\right\|\right) \cdot \|x - x^{k+0.5}\| \\
\overset{(3.2.4)}{\leq} \quad & (1+\tau)L_p\|x^{k+0.5} - x^k\|^p \cdot \|x - x^{k+0.5}\|.
\end{aligned}$$

Take $x = x^*$ and use the strong monotonicity of F:

$$\begin{aligned}
& \mu\|x^{k+0.5} - x^*\|^2 + \langle F(x^*), x^{k+0.5} - x^*\rangle \leq \langle F(x^{k+0.5}), x^{k+0.5} - x^*\rangle \\
\leq \quad & (1+\tau)L_p\|x^{k+0.5} - x^k\|^p \cdot \|x^* - x^{k+0.5}\|,
\end{aligned}$$

then we have

$$\|x^{k+0.5} - x^*\| \leq \frac{(1+\tau)L_p}{\mu}\|x^{k+0.5} - x^k\|^p. \tag{3.3.20}$$

Now, by the same analysis from (3.2.8)-(3.2.11) and take $x = x^*$, we have:

$$\langle F(x^{k+0.5}), x^{k+0.5} - x^*\rangle + \frac{\gamma_k}{2}\left(1 - \tau^2\right)\|x^{k+0.5} - x^k\|^2 \leq \frac{\gamma_k}{2}\left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right].$$

Noticing that $\langle F(x^{k+0.5}), x^{k+0.5} - x^*\rangle \geq \langle F(x^*), x^{k+0.5} - x^*\rangle \geq 0$, the above inequality implies

$$(1 - \tau^2)\|x^{k+0.5} - x^k\|^2 \leq \|x^k - x^*\|^2. \tag{3.3.21}$$

Combining (3.3.20) and (3.3.21) gives the pth*-order superlinear convergence*:

$$\|x^{k+0.5} - x^*\| \leq \frac{(1+\tau)L_p}{\mu(1-\tau^2)^{\frac{p}{2}}}\|x^k - x^*\|^p. \tag{3.3.22}$$

Note, however, that the inequality (3.3.22) only holds within each iteration and x^k in general is not converging towards x^* p^{th}-order superlinerly if a subsequent extra-gradient update is performed as in (3.2.6). In fact, once the local convergence behavior is observed, the extra-gradient update should be suppressed and the algorithm should accept $x^{k+0.5}$ as the next iterate. We shall denote

$$x^{k+1} := x^{k+0.5} \quad := \quad \text{VI}_{\mathcal{X}}\left(\tilde{F}(x; x^k) + L_p\|x - x^k\|^{p-1}(x - x^k)\right) \tag{3.3.23}$$

Algorithm 1 ARE-Restart with Local Superlinear Convergence

Require: $x^1 \in \mathcal{X}$, $0 < \alpha < 1$, $D \geq \|x^1 - x^*\|$, an inner iteration number

$$N = \left\lceil \left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}} \left(\frac{D^2}{1-\tau^2}\right)^{\frac{p-1}{p+1}} \right\rceil.$$

1: **Step 0:** Set $k := 1$.

2: **Step 1:** Let

$$x^{k+0.5} := \mathrm{VI}_{\mathcal{X}}\left(\tilde{F}(x; x^k) + L_p\|x - x^k\|^{p-1}(x - x^k)\right).$$

If

$$\|x^{k+0.5} - x^k\|^{p-1} \leq \frac{\alpha\sqrt{1-\tau^2}}{1+\tau}\frac{\mu}{L_p} \tag{3.3.24}$$

then $x^{k+1} := x^{k+0.5}$ (AR update), set $k := k+1$, and return to **Step 1**. Otherwise, go to **Step 2**.

3: **Step 2:** Let

$$x^{k+1} := \arg\min_{x\in\mathcal{X}} \langle F(x^{k+0.5}), x - x^k\rangle + \frac{L_p\|x^{k+0.5} - x^k\|^{p-1}}{2}\|x - x^k\|^2. \quad \text{(ARE-update)}$$

If $k = N$, let

$$\Gamma_N := \sum_{k=1}^{N} \frac{1}{\gamma_k}, \text{ and } \bar{x}_N := \frac{\sum_{k=1}^{N} \frac{1}{\gamma_k} x^{k+0.5}}{\Gamma_N},$$

set $x^1 := \bar{x}_N$, and return to **Step 0**. Otherwise, set $k := k+1$ and return to **Step 1**.

as *Approximation-based Regularized (AR)* update. Algorithm 1 incorporates the above decision process into ARE-restart proposed in 3.2.2, such that both the improved global iteration complexity (3.2.19) and the local p^{th}-order superlinear convergence are attained.

To verify the local convergence of Algorithm 1, we are left to show that once condition (3.3.24) is satisfied and AR update (i.e. $x^{k+1} := x^{k+0.5}$) is accepted in Step 1, the algorithm will continue repeating Step 1 to obtain

$$\|x^{k+1} - x^*\| \leq \frac{(1+\tau)L_p}{\mu(1-\tau^2)^{\frac{p}{2}}}\|x^k - x^*\|^p.$$

Indeed, from the previous analysis with $x^{k+0.5}$ replaced with x^{k+1} in (3.3.20) and (3.3.21), we have

$$\begin{aligned} \|x^{k+1} - x^k\| &\leq \frac{1}{\sqrt{1-\tau^2}}\|x^k - x^*\|, \\ \|x^{k+1} - x^*\| &\leq (1+\tau)\frac{L_p}{\mu}\|x^{k+1} - x^k\|^p, \end{aligned}$$

which implies

$$\|x^{k+2} - x^{k+1}\| \leq \frac{1}{\sqrt{1-\tau^2}} \|x^{k+1} - x^*\| \leq \frac{1+\tau}{\sqrt{1-\tau^2}} \frac{L_p}{\mu} \|x^{k+1} - x^k\|^p \leq \alpha \|x^{k+1} - x^k\|,$$

where the last inequality holds due to the condition (3.3.24). Therefore, $\{\|x^{k+1} - x^k\|\}$ becomes a contracting sequence once AR update is accepted, and Algorithm 1 will repeat Step 1 until the designated total iteration number.

We summarize the iteration complexity of Algorithm 1 in the next theorem.

Theorem 3.3.1. *Let $\{x^k\}_{k\geq 1}$ and $\{x^{k+0.5}\}_{k\geq 1}$ be generated by Algorithm 1 and suppose $F(\cdot)$ is strongly monotone with $\mu > 0$ and $\tilde{F}(x; x^k)$ is such that (3.2.4) holds with $p > 1$. The total iteration complexity to reach $\|x^k - x^*\| \leq \epsilon$ for some $\epsilon > 0$ is given by:*

$$\tilde{\mathcal{O}}\left(\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}} (D^2)^{\frac{p-1}{p+1}} + \log_p \log_2 \frac{1}{\epsilon}\right). \tag{3.3.25}$$

Proof. We omit the proof of the local iteration complexity $\log_p \log_2 \frac{1}{\epsilon}$ in view of the earlier arguments. Note that we have used $\tilde{\mathcal{O}}$ to suppress the logarithmic part in the first term of (3.3.25). Since Algorithm 1 adopts ARE-restart as global iterations, by Theorem 3.2.2, the iteration complexity requires to reach $\|x^k - x^*\|^2 \leq \hat{\epsilon}$ for some $\hat{\epsilon} > 0$ is

$$\mathcal{O}\left(\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}} (D^2)^{\frac{p-1}{p+1}} \log_{\frac{p+1}{2}} \log_2 \frac{D_0^2}{\hat{\epsilon}^2}\right).$$

In view of (3.3.21), let $\hat{\epsilon} := (1-\tau^2)\left(\frac{\alpha\sqrt{1-\tau^2}}{1+\tau} \cdot \frac{\mu}{L_p}\right)^{\frac{2}{p-1}}$, then we have:

$$\|x^{k+0.5} - x^k\|^{p-1} \leq \left(\frac{\|x^k - x^*\|^2}{1-\tau^2}\right)^{\frac{p-1}{2}} \leq \left(\frac{\hat{\epsilon}}{1-\tau^2}\right)^{\frac{p-1}{2}} \leq \frac{\alpha\sqrt{1-\tau^2}}{1+\tau} \frac{\mu}{L_p}.$$

□

As far as we know, the results on local superlinear convergence for higher-order VI methods are still quite limited in the literature. In [25], the authors establish quadratic convergence for strongly-convex-strongly-concave saddle point problem with a cubic regularized method CRN-SPP ($p = 2$). Such local quadratic convergence result is also adopted by [73] for the general p^{th}-order method. This rate is further improved to $\frac{p+1}{2}$ in [33, 53]. However, we show that for the pth-order ARE with $p > 1$, the local superlinear convergence is of the order p, achieved by the AR update (3.3.23). As shown in Algorithm 1, there is an implementable criterion (3.3.24) to determine whether to reject the extra step and continue

to converge superlinearly. We also note the difference between the results in Theorem 3.2.2 and Theorem 3.3.1. In Theorem 3.2.2, the iterates converge superlinerly after each *epoch*, which still requires $\mathcal{O}\left(\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}}(D)^{\frac{p-1}{p+1}}\right)$ inner iterations between restarts. On the other hand, when (3.3.24) is satisfied and the algorithm starts to perform only the AR updates, the iterates start to converge superlinearly after each *iteration*. The overall iteration complexity is then given in (3.3.25).

3.4 Solving Regularized VI Subproblem with $p = 2$

In the previous sections, we have presented global and local iteration complexity analysis for ARE (3.2.6). We show in Theorem 3.2.1 that the proposed simple update form of ARE guarantees the same order of improved iteration complexity as [33, 1, 53] for $p > 1$ under the monotone case, which is also optimal due to the lower bound established in [53]. For strongly monotone VI, we show that by restarting ARE, the iterates after each epochs converge at a superlinear rate with per-epoch cost $\left(\frac{L_p}{\mu}\right)^{\frac{2}{p+1}}(D^2)^{\frac{p-1}{p+2}}$. We further show that by imposing an additional condition (3.3.24) before performing the extra-step update, the local p^{th}-order superlinear convergence is guaranteed.

The aforementioned results are derived based on the assumption that the first step of ARE, which involves solving an approximation regularized VI subproblem

$$x^{k+0.5} := \mathrm{VI}_{\mathcal{X}}\left(\tilde{F}(x;x^k) + L_p\|x-x^k\|^{p-1}(x-x^k)\right), \tag{3.4.26}$$

can be efficiently performed with incomparable cost to the overall iterations. While this assumption is commonly made for higher-order methods especially for $p \geq 2$ in order to focus on analyzing the iteration complexities [33, 1, 53], we shall devote this section to the discussion on certain details for solving such subproblem (3.4.26). The rest of the section will focus on the case when $p = 2$ and the approximation mapping $\tilde{F}(x;x^k)$ is the corresponding Taylor approximation of $F(x)$, which is arguably most practical in the higher-order regime and admits some meaningful simplification and/or transformation of the subproblem.

Let us rewrite the subproblem (3.4.26) in the following form:

(3.4.27)

$$\begin{aligned}
&\text{(S)} \quad x^{k+0.5} = \mathrm{VI}_{\mathcal{X}}\left(F(x^k) + \nabla F(x^k)(x-x^k) + L_2\|x-x^k\|(x-x^k)\right)\\
&\iff \quad \text{find } x^{k+0.5} \text{ s.t.}\\
&\qquad \langle F(x^k) + \nabla F(x^k)(x^{k+0.5}-x^k) + L_2\|x^{k+0.5}-x^k\|(x^{k+0.5}-x^k), x - x^{k+0.5}\rangle \geq 0,
\end{aligned}$$

for all $x \in \mathcal{X}$. We shall refer to (3.4.27) as subproblem (S) and discuss two types of methods for solving it. In essence, the two types of methods both reduce the original subproblem (S) to another subproblem that can be more easily solved, and by solving the latter subproblem iteratively, we are able to obtain a(n) (approximated) solution to (S).

3.4.1 Reduction to VI subproblem with linear mapping

In view of (3.4.27), the operator in VI subproblem (S) takes the form of the sum of a linear operator $F(x^k) + \nabla F(x^k)(x - x^k)$ and a non-linear operator $L_2\|x - x^k\|(x - x^k)$, where the latter is the gradient mapping of the cubic regularization term and in general makes the original VI problem difficult to solve efficiently. The first type of methods then aim to reduce (3.4.27) to an easier VI problem with linear operator only, by *parameterizing* the solution $x^{k+0.5}$ as following:

$$\text{(3.4.28)} \quad \text{(SS1)} \quad x^{k+0.5}(\lambda) := \mathrm{VI}_{\mathcal{X}}\left(F(x^k) + \nabla F(x^k)(x - x^k) + \lambda(x - x^k)\right),$$

with the goal of finding

$$\text{(3.4.29)} \qquad \lambda = L_2\|x^{k+0.5}(\lambda) - x^k\|.$$

Note that subproblem (SS1) given in (3.4.28) is now a VI with linear operator $\nabla F(x^k) + \lambda I$. In particular, when $\mathcal{X} = \mathbb{R}^n$ (i.e. unconstrained), $x^{k+0.5}(\lambda)$ admits the closed-form expression:

$$\begin{aligned} & F(x^k) + \left(\nabla F(x^k) + \lambda I\right)(x^{k+0.5}(\lambda) - x^k) = 0, \\ \iff \quad & x^{k+0.5}(\lambda) = x^k - \left(\nabla F(x^k) + \lambda I\right)^{-1} F(x^k). \end{aligned}$$

This enables us to solve (3.4.29) using the next equation system with one-dimensional variable λ via Newton method:

$$f(\lambda) := \lambda^2 - L_2^2\|x^{k+0.5}(\lambda) - x^k\|^2 = 0.$$

In each iteration of the Newton method, it is then required to calculate the Jacobian of $f(\lambda)$. We shall omit the implementation details here and refer the interested readers to Section 4 in [25], which describes a more involved decomposition for calculating the Jacobian under the saddle-point problem setting (where $F(x^k)$ is the gradient descent ascent field of the saddle function).

For the case where (3.4.28) is constrained with general closed convex set, [61] proposed a *bisection* procedure for solving (3.4.29). Similar ideas are also adopted in the bisection

subroutine in [8, 33]. We briefly summarize the underlying concept of such method and refer the interested readers to [61] for analysis and implementation details. Instead of solving the equality constraint (3.4.29), one can extend it to inequality constraints:

$$\frac{L_2}{2}\|x^{k+0.5}(\lambda) - x^k\| \leq \lambda \leq 2L_2\|x^{k+0.5}(\lambda) - x^k\|. \tag{3.4.30}$$

Note a similar constraint in [8] for $p = 2$. It is shown in [61] that λ satisfying constraint (3.4.30) lies in a closed interval: $\lambda \in [t_-, t_+]$ for some $t_+ > t_- > 0$, whose range $[t_-, t_+] \subset [\alpha_-, \alpha_+]$ can be determined through solving (SS1) (3.4.28) once with initialized λ_0. Using a bisection method which uses $\lambda_+ = \sqrt{\alpha_-\alpha_+}$, the total complexity of solving (SS1) (3.4.28) with (3.4.30) being satisfied is given by:

$$\log\left(\frac{\log(\alpha_+/\alpha_-)}{\log(t_+/t_-)}\right).$$

Therefore, solving the subproblem (S) boils down to how to solve (SS1) efficiently at each iteration of the bisection method proposed in [61]. Since the VI operator for solving (SS1) is linear, if the constraint set $\mathcal{X}$ takes simpler forms such as $\mathcal{X} := \mathbb{R}^n_+$, (3.4.28) can be reduced to a *linear complementarity problem* (LCP), which then can be solved efficiently by using, for example, interior point method [19].

3.4.2 Reduction to gradient projection

In the second type of method, we propose an alternative procedure to solve (S), which applies a first-order iterative method (an inner loop) to solve the VI problem (3.4.27) for an approximated solution for $x^{k+0.5}$. Let us define the following operator

$$F'(x; x^k) := F(x^k) + \nabla F(x^k)(x - x^k) + L_2\|x - x^k\|(x - x^k). \tag{3.4.31}$$

A naive way to implement the inner loop is to directly apply, for example, the extra-gradient method to solve $\mathrm{VI}_{\mathcal{X}}(F'(x; x^k))$. The potential issue lies in the fact that $F'(x; x^k)$ is in general not Lipschitz continuous over the whole constraint $\mathcal{X}$ due to the mapping $L_2\|x - x^k\|(x - x^k)$. Therefore, no guarantee on the performance of the inner loop can be established.

In order to implement a more efficient procedure for solving (3.4.27) (or succinctly $\mathrm{VI}_{\mathcal{X}}(F'(x; x^k))$), we first discuss a specific instance of the ARE update introduced in Section 3.2 with $p = 1$. Let us define $F'(x) := F'(x; x^k)$ to simplify the notation and note that $F'(x)$ takes the summation form $F'(x) = H(x) + G(x)$. Assume that $H(\cdot)$ is Lipschitz continuous with constant L_H and strongly monotone with modulus μ_H, and $G(\cdot)$ is monotone. Consider the approximation operator $\tilde{F}'(x; y) := H(y) + G(x)$, and we have:

$$\left\|\tilde{F}'(x; y) - F(x)\right\| = \|H(x) - H(y)\| \leq L_H\|x - y\|.$$

Therefore, based on the results in Section 3.2 (in particular Theorem 3.2.2), the following update procedure

$$(3.4.32)\begin{cases} \bar{x}^{t+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(H(\bar{x}^t)+G(x)+L_H(x-\bar{x}^t)\right), \\ \\ \bar{x}^{t+1} & := & \arg\min\limits_{x\in\mathcal{X}} \langle H(\bar{x}^{t+0.5})+G(\bar{x}^{t+0.5}), x-\bar{x}^t\rangle + \frac{L_H}{2}\|x-\bar{x}^t\|^2, \end{cases}$$

for $t = 0, 1, 2, ...$ is guaranteed to converge to an $\bar{\epsilon}$ solution with iteration complexity $\mathcal{O}\left(\frac{L_H}{\mu_H}\log\frac{1}{\bar{\epsilon}}\right)$. Indeed, method (3.4.32) is nothing but an ARE update instance (3.2.6), where $F(x) := F'(x) = H(x) + G(x)$ and $\tilde{F}(x;\bar{x}^t) := \tilde{F}'(x;\bar{x}^t) = H(\bar{x}^t) + G(x)$, and $p = 1$. Let us denote $H(x) = F(x^k) + \nabla F(x^k)(x - x^k)$ and $G(x) = L_2\|x - x^k\|(x - x^k)$. Indeed, $H(x)$ is Lipschitz continuous with $L_H = \|\nabla F(x^k)\| \le L$, and $H(x) + G(x)$ is strongly monotone with $\mu_H = \mu > 0$ provided the original operator $F(x)$ is strongly monotone with $\mu > 0$. Under this formulation, solving subproblem (S) (3.4.27) is equivalent to solving $\mathrm{VI}_{\mathcal{X}}(F'(x)) = \mathrm{VI}_{\mathcal{X}}(H(x) + G(x))$ and can be solved approximately by the iterative procedure (3.4.32) with iteration complexity $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\bar{\epsilon}}\right)$.

We now show that each iteration of (3.4.32) can be further reduced to two gradient projection steps, whose computational cost is significantly reduced compared to directly solving the original VI subproblem (S). Note that the second step for updating $\bar{x}^{t+1}$ requires a gradient projection step, while the first step requires solving $\bar{x}^{t+0.5}$ from a VI problem in the following form:

$$\langle H(\bar{x}^t) + L_2\|\bar{x}^{t+0.5} - x^k\|(\bar{x}^{t+0.5} - x^k) + L_H(\bar{x}^{t+0.5} - \bar{x}^t), x - \bar{x}^{t+0.5}\rangle \ge 0, \quad \forall x \in \mathcal{X},$$

which is optimality condition of the optimization problem:

$$(3.4.33) \qquad \min_{x\in\mathcal{X}} \quad \frac{L_2}{3}\|x - x^k\|^3 + \frac{L_H}{2}\|x - \bar{x}^t\|^2 + H(\bar{x}^t)^\top x.$$

The following analysis adopts a similar reformulation as proposed in [67] to solve (3.4.33) . Let us first reformulate (3.4.33) into

$$\begin{aligned} & \arg\min_{x\in\mathcal{X}} \quad \frac{L_2}{3}\|x - x^k\|^3 + \frac{L_H}{2}\|x - \bar{x}^t\|^2 + H(\bar{x}^t)^\top x \\ = & \arg\min_{x\in\mathcal{X}} \quad \frac{L_2}{3}\|x - x^k\|^3 + \frac{L_H}{2}\left(\|x - x^k\|^2 + \|x - \bar{x}^t\|^2 - \|x - x^k\|^2\right) + H(\bar{x}^t)^\top x \\ = & \arg\min_{x\in\mathcal{X}} \quad \frac{L_2}{3}\|x - x^k\|^3 + \frac{L_H}{2}\|x - x^k\|^2 + L_H(x^k - \bar{x}^t)^\top x + H(\bar{x}^t)^\top x \\ = & \arg\min_{x\in\mathcal{X}} \quad \frac{L_2}{3}\|x - x^k\|^3 + \frac{L_H}{2}\|x - x^k\|^2 + \left(L_H(x^k - \bar{x}^t) + H(\bar{x}^t)\right)^\top (x - x^k). \end{aligned}$$

Denote

$$g_t(x^k) = L_H(x^k - \bar{x}^t) + H(\bar{x}^t) = F(x^k) + \left(\nabla F(x^k) - I\right)(\bar{x}^t - x^k),$$

for a given fixed x^k. Since

$$\frac{1}{3}r^3 = \max_{\tau \geq 0} \quad r^2\tau - \frac{2}{3}\tau^{\frac{3}{2}},$$

we have

$$\begin{aligned}
& \min_{x\in\mathcal{X}} \quad \frac{L_2}{3}\|x-x^k\|^3 + \frac{L_H}{2}\|x-x^k\|^2 + g_t(x^k)^\top(x-x^k) \\
= & \min_{x\in\mathcal{X}} \max_{\tau\geq 0} \quad L_2\left(\tau\|x-x^k\|^2 - \frac{2}{3}\tau^{\frac{3}{2}}\right) + \frac{L_H}{2}\|x-x^k\|^2 + g_t(x^k)^\top(x-x^k) \\
= & \max_{\tau\geq 0}\left(-\frac{2}{3}L_2\tau^{\frac{3}{2}} + \min_{x\in\mathcal{X}}\left\{g_t(x^k)^\top(x-x^k) + \left(L_2\tau + \frac{L_H}{2}\right)\|x-x^k\|^2\right\}\right).
\end{aligned} \tag{3.4.34}$$

The inner minimization gives a closed-form solution, denoted as

$$\bar{x}^{t+0.5}(\tau) = \mathbf{P}_{\mathcal{X}}\left(x^k - \frac{1}{2L_2\tau + L_H}g_t(x^k)\right),$$

where $\mathbf{P}_{\mathcal{X}}$ is the projection operator onto $\mathcal{X}$. On the other hand, the outer maximization of (3.4.34) is a simple one-dimensional concave maximization, and the solution given the expression of $\bar{x}^{t+0.5}(\tau)$ is given by the following:

$$\tau^* = \arg\max_{\tau} \quad -\frac{2}{3}L_2\tau^{\frac{3}{2}} - \frac{1}{4L_2\tau + 2L_H}\|g_t(x^k)\|^2,$$

which can be solved efficiently by any common software.

To summarize, the update process in (3.4.32) can be rewritten into

$$\text{(SS2)} \quad \begin{cases} \tau^* := \arg\max_{\tau} \quad -\frac{2}{3}L_2\tau^{\frac{3}{2}} - \frac{1}{4L_2\tau+2L_H}\|g_t(x^k)\|^2, \\ \bar{x}^{t+0.5} = \bar{x}^{t+0.5}(\tau^*) := \mathbf{P}_{\mathcal{X}}\left(x^k - \frac{1}{2L_2\tau^*+L_H}g_t(x^k)\right), \\ \bar{x}^{t+1} := \arg\min_{x\in\mathcal{X}} \langle H(\bar{x}^{t+0.5}) + G(\bar{x}^{t+0.5}), x - \bar{x}^t\rangle + \frac{L_H}{2}\|x-\bar{x}^t\|^2, \end{cases} \tag{3.4.35}$$

whose major cost lies in performing two gradient projection steps. Process (3.4.35) is then preformed iteratively until we obtain an approximate solution $\|\bar{x}^t - x^{k+0.5}\| \leq \bar{\epsilon}$ with iteration complexity $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\bar{\epsilon}}\right)$. Compared to the methods discussed in Section 3.4.1, the method discussed in this section in general can require more inner iterations to operate with, but at the same time solving subproblem (SS2) can also be performed with much less cost than solving (SS1).

3.5 Structured ARE Schemes

In the previous sections, we analyze the convergence properties of ARE in its general update form (3.2.6) without specifying the approximation mapping $\tilde{F}(x; x^k)$ and the corresponding order p. Such general form is powerful in that it enables us to establish unified analysis for potentially many different specific methods. We devote this section to the discussion on some of these examples and the connections to existing methods in the literature.

Consider the following structured VI where the operator is given in the *composite form*:

$$F(x) = H(x) + G(x). \tag{3.5.36}$$

We shall discuss different realizations of ARE in solving $\mathrm{VI}_{\mathcal{X}}(F(x))$. The first immediate example is the extra-gradient method, which is equivalent to ARE with $p = 1$ and $\tilde{F}(x; x^k) := F(x^k)$:

$$\left\{ \begin{array}{rcl} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(F(x^k) + L_1(x - x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x \in \mathcal{X}} \langle F(x^{k+0.5}), x - x^k \rangle + \frac{L_1}{2}\|x - x^k\|^2. \end{array} \right. \tag{3.5.37}$$

The extra-gradient method (3.5.37) treats $F(x)$ as a single operator without using the specific composite structure (3.5.36). The proposed ARE, however, provides the possibilities of using alternative approximation operator $\tilde{F}(x; x^k)$ in the update. In particular, consider the case $p = 1$ and $\tilde{F}(x; x^k) := H(x^k) + G(x)$:

$$\left\{ \begin{array}{rcl} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(H(x^k) + G(x) + L_1(x - x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x \in \mathcal{X}} \langle F(x^{k+0.5}), x - x^k \rangle + \frac{L_1}{2}\|x - x^k\|^2, \end{array} \right. \tag{3.5.38}$$

where the update of $x^{k+0.5}$ can be viewed as a combined gradient-projection/proximal point step, or a *proximal gradient* (projection) update. As we have also seen in Section 3.4.2, the potential advantage of performing such update is that we are able to relax the Lipschitz continuity assumption made for the overall operator $F(x)$. Indeed, the required condition for the iteration complexity results to hold for ARE is given by (3.2.4), while in the update (3.5.38) we have:

$$\|\tilde{F}(x; x^k) - F(x)\| = \|H(x^k) - H(x)\| \leq L_H \|x - x^k\|,$$

where we assume L_H is the Lipschitz constant for $H(x)$. Therefore, only the Lipschitz conitnuity of $H(x)$ is required, and the constant L_1 in (3.5.38) can be replaced with L_H. In the example discussed in Section 3.4.2, the VI subproblem for solving $x^{k+0.5}$ can be further reduced to simpler forms if $G(x)$ is the gradient of some convex function.

We can also extend the original problem $\mathrm{VI}_{\mathcal{X}}(F(x))$ to the monotone inclusion problem:

$$0 \in H(x^*) + G(x^*),$$

where $G : \mathcal{X} \rightrightarrows \mathbb{R}^n$ is a set-valued maximal monotone operator. Well-known maximal monotone operators include ∂f, the subdifferential of a proper closed convex function f, and $N_{\mathcal{X}}(x)$, the normal cone of a closed convex set $\mathcal{X}$. For further discussion regarding maximal monotone operator and monotone inclusion problem, the interested readers are referred to [19]. In view of the monotone inclusion problem, the same ARE update discussed earlier (3.5.38) can be adjusted accordingly:

$$\text{(3.5.39)} \quad \begin{cases} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(H(x^k) + u^{k+0.5} + L_1(x - x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x\in\mathcal{X}} \langle H(x^{k+0.5}) + u^{k+0.5}, x - x^k\rangle + \frac{L_1}{2}\|x - x^k\|^2, \end{cases}$$

where $u^{k+0.5} \in G(x^{k+0.5})$. Note that in the above expression, $x^{k+0.5}$ equivalently satisfies the following equation:

$$H(x^k) + u^{k+0.5} + L_1(x^{k+0.5} - x^k) = 0.$$

Substituting $u^{k+0.5} = -H(x^k) - L_1(x^{k+0.5} - x^k)$ in the update of x^{k+1}, we get the following scheme:

$$\text{(3.5.40)} \begin{cases} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(H(x^k) + u^{k+0.5} + L_1(x - x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x\in\mathcal{X}} \langle H(x^{k+0.5}) - H(x^k), x - x^k\rangle + \frac{L_1}{2}\|x - x^{k+0.5}\|^2, \end{cases}$$

which is the modified forward-backward update proposed in [97] by noticing that $x^{k+0.5}$ in (3.5.40) is also equivalent to the forward-backward step $x^{k+0.5} = \left(I + \frac{1}{L_1}G\right)^{-1}\left(I - \frac{1}{L_1}H\right)x^k$. This shows that the modified forward-backward method is indeed another important instance of ARE for the more general monotone inclusion problem.

Moving forward to the higher-order ($p \geq 2$) ARE schemes, an immediate example is to take the Taylor approximation $\tilde{F}(x; x^k) := \sum\limits_{i=0}^{p-1} \frac{1}{i!}\nabla^i F(x^k)[x - x^k]^i$, resulting in the following update:

(3.5.41)

$$\begin{cases} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(\sum\limits_{i=0}^{p-1} \frac{1}{i!}\nabla^i F(x^k)[x - x^k]^i + L_p\|x - x^k\|^{p-1}(x - x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x\in\mathcal{X}} \langle F(x^{k+0.5}), x - x^k\rangle + \frac{L_p\|x^{k+0.5} - x^k\|^{p-1}}{2}\|x - x^k\|^2. \end{cases}$$

The above update (3.5.41) can be viewed as equivalent forms of the NPE [61] ($p = 2$) and higher-order mirror-prox [8, 1] ($p \geq 2$). In view of the previous discussion, it is

then natural to consider the higher-order approximation operator in the form $\tilde{F}(x;x^k) := \sum_{i=0}^{p-1} \frac{1}{i!}\nabla^i H(x^k)[x-x^k]^i + G(x)$ for the specific composite structure (3.5.36), and the next scheme follows:

(3.5.42)
$$\left\{\begin{array}{rcl} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(\sum_{i=0}^{p-1}\frac{1}{i!}\nabla^i H(x^k)[x-x^k]^i + G(x) + L_p\|x-x^k\|^{p-1}(x-x^k)\right), \\ x^{k+1} & := & \arg\min_{x\in\mathcal{X}} \langle F(x^{k+0.5}), x-x^k\rangle + \frac{L_p\|x^{k+0.5}-x^k\|^{p-1}}{2}\|x-x^k\|^2. \end{array}\right.$$

The above scheme (3.5.42) can be viewed as generalization of several existing methods. In addition to generalizing the higher-order mirror-prox method, it also generalizes the modified forward-backward method (in the form (3.5.38)) to p^{th}-order. Furthermore, it generalizes the tensor method proposed in [14] for composite optimization to solving composite VI. Indeed, consider the following problem:

$$\min_{x\in\mathcal{X}} \quad h(x) + g(x).$$

To simplify the discussion, assume both h, g are convex and differentiable and denote $H(x) := \nabla h(x)$ and $G(x) := \nabla g(x)$. The following inequality defines the solution $x^{k+0.5}$ in (3.5.42):

$$\begin{array}{rl} & \langle \sum_{i=0}^{p-1}\frac{1}{i!}\nabla^i H(x^k)[x^{k+0.5}-x^k]^i + G(x^{k+0.5}) + L_p\|x^{k+0.5}-x^k\|^{p-1}(x^{k+0.5}-x^k), x-x^{k+0.5}\rangle \\ \geq & 0, \\ & \forall x\in\mathcal{X}, \end{array}$$

which is equivalent to

$$x^{k+0.5} := \arg\min_{x\in\mathcal{X}} \sum_{i=1}^{p}\frac{1}{i!}\nabla^i h(x^k)[x-x^k]^i + g(x) + \frac{L_p}{p+1}\|x-x^k\|^{p+1}. \tag{3.5.43}$$

Note that in the context of composite optimization [14], the problem is unconstrained and the minimization step (3.5.43) is performed over the domain of $g(x)$. In addition, while the acceleration in optimization requires an additional sequence $\{y^k\}$ so that the update (3.5.43) is performed at y^k instead of x^k (for example, FISTA [6] and Nesterov's accelerated tensor method [69]), the acceleration in VI in general takes the form of the extra-gradient step such as the update of x^{k+1} in (3.5.42).

Next, we further consider the following more general composite VI model with the operator:

$$F(x) = H(G(x)). \tag{3.5.44}$$

Obviously, if we let $G(x) = G_1(x) + G_2(x)$ and $H(x) = x$, the general composite operator (3.5.44) reduces to the special case in the summation form (3.5.36). This general model enables us to extend the approximation schemes discussed earlier, as shown in the next two examples. The first example is an *outer approximation*:

$$\tilde{F}(x; x^k) := \tilde{H}(G(x); G(x^k)),$$

which replaces the outer operator $H(\cdot)$ with an approximation operator $\tilde{H}(\cdot\,; y)$ that satisfies the condition (3.2.4) with some fixed y and constant $L_p := L_H$. The resulting overall approximation $\tilde{F}(x; x^k)$ hence satisfies (3.2.4) as well from the following bound:

$$\begin{aligned} \left\|\tilde{F}(x; x^k) - F(x)\right\| &= \left\|\tilde{H}(G(x); G(x^k)) - H(G(x))\right\| \\ &\leq \tau L_H \left\|G(x) - G(x^k)\right\|^p \leq \tau L_H L_G^p \|x - x^k\|^p, \end{aligned}$$

where we also assume $G(x)$ is Lipschitz continuous with constant L_G. As an exemplifying scheme, let $\tilde{H}(\cdot\,; y)$ be the Taylor approximation of $H(\cdot)$ with $p = 2$, which results in the next update scheme:

(3.5.45)
$$\left\{ \begin{array}{lcl} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}} \left(H(G(x^k)) + \nabla H(G(x^k)) \left(G(x) - G(x^k)\right) + L_H L_G^2 \|x - x^k\| (x - x^k) \right), \\ x^{k+1} & := & \arg\min\limits_{x \in \mathcal{X}} \langle F(x^{k+0.5}), x - x^k \rangle + \frac{L_H L_G^2 \|x^{k+0.5} - x^k\|}{2} \|x - x^k\|^2. \end{array} \right.$$

Note that in this example, even if the outer approximation operator $\tilde{H}(\cdot\,; y)$ is the Taylor approximation, the overall approximation operator $\tilde{F}(x; x^k) := H(G(x^k)) + \nabla H(G(x^k)) \left(G(x) - G(x^k)\right)$ is not (it is not even linear unless $G(\cdot)$ is). In general, $\tilde{H}(\cdot\,; y)$ needs not be the Taylor approximation but can be any approximation satisfying (3.2.4), such as the ones discussed earlier.

The second example based on the composite VI model (3.5.44) is an *inner approximation*:

$$\tilde{F}(x; x^k) := H(\tilde{G}(x; x^k)),$$

which replaces the inner operator $G(x)$ with an approximation operator $\tilde{G}(x; x^k)$ that satisfies the condition (3.2.4), with the constant now defined as $L_p := L_G$. Similarly, we have:

$$\begin{aligned} \left\|\tilde{F}(x; x^k) - F(x)\right\| &= \left\|H(\tilde{G}(x; x^k)) - H(G(x))\right\| \\ &\leq L_H \left\|\tilde{G}(x; x^k) - G(x)\right\| \leq \tau L_H L_G \|x - x^k\|^p, \end{aligned}$$

which indicates that $\tilde{F}(x; x^k)$ also satisfies (3.2.4) as long as $H(\cdot)$ is Lipschitz continuous with L_H. If $\tilde{G}(x; x^k)$ is the Taylor approximation of $G(x)$ at x^k with $p = 2$, we have the

following scheme:

(3.5.46)
$$\begin{cases} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(H\left(G(x^k)+\nabla G(x^k)(x-x^k)\right)+L_H L_G\|x-x^k\|(x-x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x\in\mathcal{X}} \langle F(x^{k+0.5}), x-x^k\rangle + \frac{L_H L_G\|x^{k+0.5}-x^k\|}{2}\|x-x^k\|^2. \end{cases}$$

Again, the overall approximation operator $\tilde{F}(x;x^k)$ is not Taylor approximation even if the inner approximation operator $\tilde{G}(x;x^k)$ is, and it is not linear unless $H(\cdot)$ is. While the examples in (3.5.45) and (3.5.46) use a similar concept to construct the approximation operator $\tilde{F}(x;x^k)$, the resulting update scheme can be quite different given how we identify the specific composite structure $F(x)=H(G(x))$ in a problem.

In this section, we first discuss several structured ARE schemes based on the composite form of the operator (3.5.36), which either coincides with or generalizes existing methods. We further discuss the more general composite VI model (3.5.44) and present two different examples to illustrate the concept of outer approximation and inner approximation. We remark that composite VI may take even more general forms such as multiple layers $F(x)=H_1(H_2(...(H_n(x))))$, or multiple blocks $F(x)=H(G_1(x),G_2(x),...,G_n(x))$, or arbitrary combinations of these two. Developing specific schemes based on these composite forms can be highly dependent on each individual problem at hand and the subproblem of solving $x^{k+0.5}$ may be difficult. However, the purpose of this work is to reveal the potentials of a general scheme of approximation used in the ARE framework, by pointing out possibilities other than the most commonly applied Taylor approximations in many existing schemes. By taking the structure of the VI operator into consideration, ARE can possibly include even more complicated schemes than the ones discussed in this section. As long as certain assumptions are satisfied and one is able to develop efficient subroutines for solving the VI subproblem, the results established in earlier sections can immediately provide optimal iteration complexity guarantee for the new scheme.

3.6 Numerical Experiments

In this section, we examine the convergence of ARE and ARE-restart with $p=2$ and compare the performance with other common first-order methods. We consider the following

unconstrained saddle point problem in the experiment:

$$\begin{aligned}\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} f(x,y) &= \frac{1}{M_1}\sum_{i=1}^{M_1}\ln(1+e^{-a_i^\top x})+\frac{\lambda}{2}\|x\|^2 \\ &\quad +x^\top Ay-\frac{1}{M_2}\sum_{j=1}^{M_2}\ln(1+e^{-b_j^\top y})-\frac{\lambda}{2}\|y\|^2. \end{aligned} \tag{3.6.47}$$

To transform the saddle point problem (3.6.47) into equivalent VI formulation, let us redefine the VI variable as $u=(x,y)^\top$ and the operator

$$F(u)=\begin{pmatrix} -\frac{1}{M_1}\sum_{i=1}^{M_1}\frac{a_i}{1+e^{-a_i^\top x}}+\lambda x+Ay \\ -\frac{1}{M_2}\sum_{j=1}^{M_2}\frac{b_j}{1+e^{-b_j^\top y}}+\lambda y-A^\top x \end{pmatrix},$$

with the problem defined as $F(u)=0$. The ARE and ARE-restart implemented in the experiment specifically use the Taylor approximation as the approximation operator $\tilde{F}(u,u^k):=F(u^k)+\nabla F(u^k)(u-u^k)$ and can be expressed as:

$$\left\{\begin{array}{rcl} u^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(F(u^k)+\nabla F(u^k)(u-u^k)+L_2\|u-u^k\|(u-u^k)\right), \\ u^{k+1} & := & \arg\min_{u\in\mathcal{X}}\langle F(u^{k+0.5}),u-u^k\rangle+\frac{L_2\|u^{k+0.5}-u^k\|}{2}\|u-u^k\|^2. \end{array}\right. \tag{3.6.48}$$

Since the original saddle point problem is unconstrained, we have $\mathcal{X}:=\mathbb{R}^n\times\mathbb{R}^m$, and the VI subproblem for solving $u^{k+0.5}$ is equivalent to solving the equation:

$$F(u^k)+\nabla F(u^k)(u^{k+0.5}-u^k)+L_2\|u^{k+0.5}-u^k\|(u^{k+0.5}-u^k)=0,$$

which can be solved via a Newton method (see discussions in Section 3.4.1). For a more detailed implementation, the interested readers are referred to Section 4 in [25]. The restart procedure of update (3.6.48) is described in Section 3.2.2, and we use a pre-defined number for inner iterations between each restart.

The experiment is conducted under Matlab 2018 environment, and the problem parameters are as follows. The number of date points is $M_1=M_2=100$; the problem dimensions are $m=2n=50$; the elements of a_i,b_j,A are generated by independent standard normal distribution; the second-order smoothness constant L_2 is estimated as 0.3. Note that the operator $F(u)$ is strongly monotone with modulus λ, which is varied to observe different convergence behaviors. The purpose of the experiments is to verify the convergence of ARE and ARE-restart, and we use the first-order methods, extra-gradient and OGDA, as the benchmarks for comparison. The convergence is measured as $\|F(u)\|$, and the results are presented in Figure 3.1-3.3.

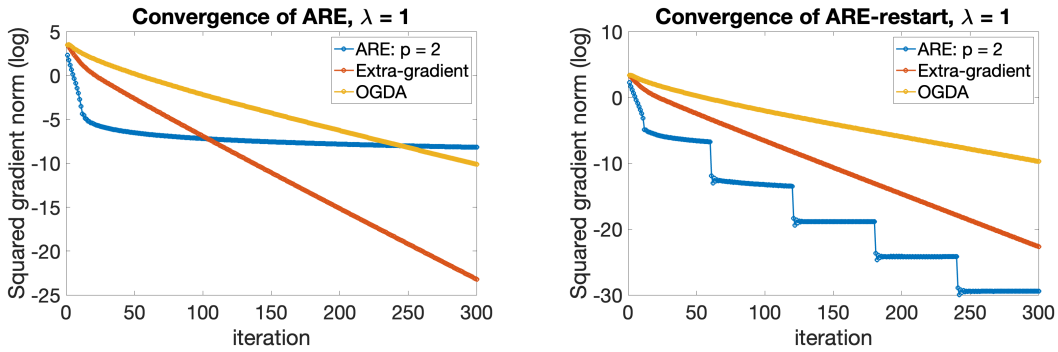

Figure 3.1: Convergence in strongly monotone VI with $\lambda = 1$

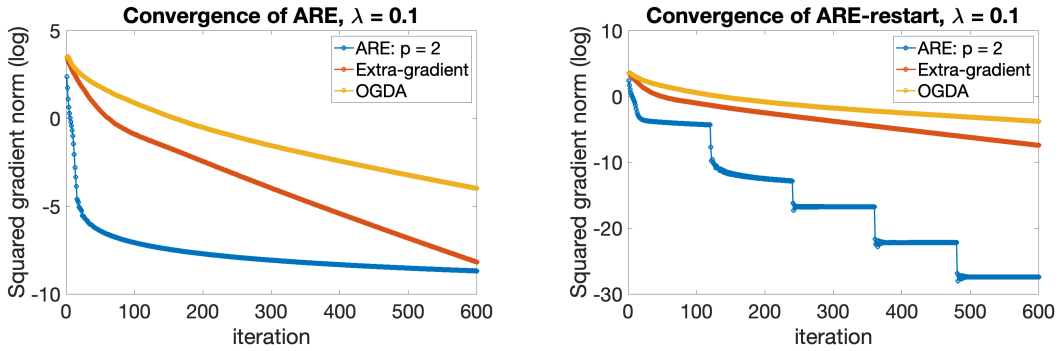

Figure 3.2: Convergence in strongly monotone VI with $\lambda = 0.1$

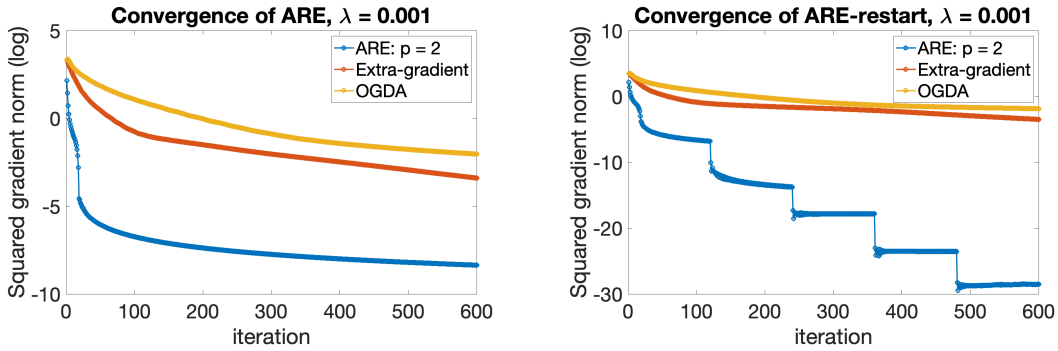

Figure 3.3: Convergence in strongly monotone VI with $\lambda = 0.001$

The convergence of ARE are shown in the left plots of Figure 3.1, Figure 3.2, and Figure 3.3. All of them show clear sublinear convergence for the averaged iterates

$$\bar{u}_k := \frac{\sum_{i=1}^{k} \frac{u^{i+0.5}}{\gamma_i}}{\Gamma_k}, \quad \Gamma_k := \sum_{i=1}^{k} \gamma_i^{-1},$$

where $\gamma_i = L_2 \|u^{i+0.5} - u^i\|$. Indeed, a sublinear convergence rate $\mathcal{O}\left(1/k^{\frac{3}{2}}\right)$ is guaranteed for ARE with $p = 2$. However, when the problem is strongly monotone ($\lambda = 1$), it will take significantly more iterations to converge to very high precision ($\|F(u)\| < 10^{-10}$) compared to the first-order methods extra-gradient and OGDA, which are designed to better exploit the strong monotonicity and admit linear convergence. However, when λ is small ($\lambda = 0.001$) and the problem becomes closer to a VI that is merely monotone, the performance of these first-order methods deteriorate fast to sublinear convergence that is significantly slower than ARE, a second-order method. On the other hand, ARE-restart (right plots of Figure 3.1, Figure 3.2, and Figure 3.3) shows clear improvement over the first-order methods regardless of the strong monotonicity modulus λ. The process of restart is crucial in these experiments to take advantage of the strong monotonicity in the problem and bring the convergence of ARE beyond sublinear convergence to linear, or even superlinear, convergence. In the results shown in Figure 3.1-Figure 3.3, the superlinear convergence happens immediately after each restart, followed by sublinear convergence in the rest of the epoch before the next restart. This particular convergence behavior enables the iterates of ARE-restart to quickly converge to high precision within much fewer iterations compared to ARE or other first-order methods.

3.7 Conclusions

In this chapter, we propose the approximation-based regularized extra-gradient (ARE) scheme for solving monotone VI. The key feature of ARE is to solve a regularized VI subproblem in the first step, where the operator consists of a general approximation mapping satisfying a pth-order Lipschitz bound (3.2.4) and the gradient mapping of a $(p+1)^{\text{th}}$-order regularization. Iteration complexities are established for both monotone VI (ARE) and strongly monotone VI (ARE-restart), and the results match the lower bound for general p^{th}-order methods. We further analyze the local convergence behavior for strongly monotone VI when $p > 1$ and establish p^{th}-order superlinear convergence, which is an improvement over the existing results. By introducing the general approximation mapping that satisfies the Lipschitz bound, ARE can be viewed as a more general framework that includes multiple existing methods in the literature. As a result, unified results can be es-

tablished for different methods under the general ARE framework. We then discuss detailed implementations for solving the regularized VI subproblem under special cases, as well as some specialized ARE schemes if the VI operator has a composite structure.

Chapter 4

New First-Order Algorithms for Stochastic Variational Inequality Problems

In this chapter, we propose two new solution schemes to solve the stochastic strongly monotone variational inequality problems: the stochastic extra-point solution scheme and the stochastic extra-momentum solution scheme. The first one is a general scheme based on updating the iterative sequence and an auxiliary extra-point sequence. In the case of deterministic VI model, this approach includes several state-of-the-art first-order methods as its special cases. The second scheme combines two momentum-based directions: the so-called *heavy-ball* direction and the *optimism* direction, where only one projection per iteration is required in its updating process. Both of these schemes can in fact be viewed as concrete instances of algorithms generated by following the general extra-point approach discussed in Chapter 2. We show that, if the variance of the stochastic oracle is appropriately controlled, then both schemes can be made to achieve optimal iteration complexity of $\mathcal{O}\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$ to reach an ϵ-solution for a strongly monotone VI problem with condition number κ. As a specific application to stochastic VI, we demonstrate how to incorporate a zeroth-order approach for solving stochastic minimax saddle-point problems in our schemes, where only noisy and biased samples of the objective can be obtained, with a total sample complexity of $\mathcal{O}\left(\frac{\kappa}{\epsilon}\right)$.

4.1 Introduction

Given a constraint set $\mathcal{Z} \subseteq \mathbb{R}^n$ and a mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$, the classical variational inequality problem is to find $z^* \in \mathcal{Z}$ such that

$$F(z^*)^\top (z - z^*) \geq 0, \quad \forall z \in \mathcal{Z}. \tag{4.1.1}$$

In this chapter, we consider a *stochastic* version of problem (4.1.1), where the exact evaluation of the mapping $F(\cdot)$ is inaccessible. Instead, only a *stochastic oracle* is available. The stochasticity in question may stem from, e.g., the non-deterministic nature of mixed strategies of the players in a game-setting, or simply because of the difficulty in evaluating the mapping itself. The latter has become more pronounced in the literature, due to its recent-found application as a training/learning subproblem in machine learning and/or statistical learning. The so-called *stochastic oracle* is a noisy estimation of the mapping $F(\cdot)$, and an iterative scheme that incorporates such oracle is known as *stochastic approximation* (SA). As far as we know, the first proposal to use such approach for stochastic optimization can be traced back to the seminal work of Robbins and Monro [81]. In 2008, Jiang and Xu [32] followed the SA approach to solve VI models. Since then, efforts have been made to extend existing deterministic methods to the stochastic VI models; see e.g. [35, 106, 36, 43, 31, 29, 30].

Let us start our discussion by introducing the assumptions made throughout the chapter. We consider VI model (4.1.1) where $\mathcal{Z}$ is a closed convex set. Moreover, throughout this section we assume the strong monotonicity (1.1.2) ($\mu > 0$) and Lipschitz continuity of F (1.1.3), with $\kappa := \frac{L}{\mu} \geq 1$ denoted as the condition number of (4.1.1).

The *stochastic oracle* of the mapping, denote by $\hat{F}(z, \xi)$, takes a random sample $\xi \in \Xi$ from some sample space Ξ. The oracle is required to satisfy:

$$\|\mathbb{E}[\hat{F}(z, \xi)] - F(z)\| \leq \delta, \quad \mathbb{E}\left[\|\hat{F}(z, \xi) - F(z)\|^2\right] \leq \sigma^2 \tag{4.1.2}$$

for all $z \in \mathcal{Z}$, where $\delta, \sigma \geq 0$ are some constants. In other words, we assume that both the bias and the deviation are uniformly upper-bounded.

In this chapter, we propose two stochastic first-order schemes: the *stochastic extra-point scheme* and the *stochastic extra-momentum scheme*. The first scheme maintains two sequences of iterates featuring several well-known first-order search directions such as: the extra-gradient [41, 96], the heavy-ball [78], Nesterov's extrapolation [65, 66], and the optimism direction [79, 58]. The second scheme, on the other hand, specifically combines the *heavy-ball* momentum and the *optimism* momentum in its updating formula, and maintains

only one sequence throughout the iterations, therefore requiring only *one* projection per iteration.

The deterministic counterpart of these methods can be found in Chapter 2. In particular, they can be viewed as concrete realizations of the general extra-point approach proposed in Chapter 2 for solving strongly monotone VI. In the stochastic context, we show that as long as the variance can be reduced throughout the iterations, they yield the *optimal* iteration complexity $\mathcal{O}(\kappa \ln(1/\epsilon))$ (cf. [108]) to reach ϵ-solution: $\|z^k - z^*\|^2 \leq \epsilon$, with an additional biased term depending on δ. In Section 4.3, we demonstrate an application to the stochastic black-box minimax saddle-point problem where only noisy function values $f(x, y)$ are accessible. This application is particularly relevant, given its applications in machine learning, where the training data set may be very large and evaluating exact gradient/function value is usually impractical. Through a smoothing technique, we show how to incorporate *stochastic zeroth-order gradient* as our update directions in either the stochastic extra-point scheme or the stochastic extra-momentum scheme. We show that both approaches yield an iteration complexity of $\mathcal{O}(\kappa \ln(1/\epsilon))$ and a sample-complexity of $\mathcal{O}\left(\frac{\kappa}{\epsilon}\right)$.

In fact, while the individual components in our proposed schemes have been studied extensively as different *accelerated* first-order methods in either variational inequality formulation or optimization, there is a surge of interest in their practical performance in the machine learning community, especially in training *Generative Adversarial Networks* (GANs) [22]. Formulated as saddle-point problems (in particular a zero-sum game), GANs are known to be hard to train, and conventional methods applied to standard deep learning such as (stochastic) gradient descent or mirror descent do not work well. Daskalakis *et al.* [11] propose to use Optimistic Mirror Descent (OMD) to train GAN and show better performance over mirror descent as well as theoretical convergence for bilinear saddle point problems. Mertikopoulos *et al.* [57] study a variant of OMD, which could also be understood as an extra-gradient method. Liang and Stokes [49] study both extra-gradient method and OMD in bilinear saddle point problems. Gidel *et al.* [21] analyze the GAN formulation in a more general VI framework and compare methods that use extrapolation (extra-gradient method) and extrapolation from the past (OGDA). These results have shown that incorporating search directions beyond simple gradient information, such as using optimism (OMD) or extra-gradient, is indeed important in practical GAN training, in addition to their stronger theoretical guarantees.

With these versatile optimal first-order methods in mind, this chapter aims at a holistic study through the two proposed stochastic schemes in a general framework. Both schemes render a wider range of search directions than the existing first-order methods. On the

one hand, each of them is helpful in the practical performance compared to pure stochastic gradient method. On the other hand, there is no evidence showing that one should be preferred over another in general and it often requires multiple trials before a better method can be determined for a specific problem class. The proposed schemes thus provide a general framework of *learning*, from the data of a given class of problem instances, a better configuration among these search directions that does not necessarily cling to any specific method. Therefore, the parameters associated with each search direction could and *should* be tuned differently from problem-class to problem-class in order to secure good practical performances.

4.2 The Stochastic First-Order Methods for Strongly Monotone VI Problems

Let us start this section by introducing the notations to facilitate our analysis. We shall denote the stochastic oracle as $\hat{F}(\cdot)$, suppressing the notation ξ whenever it is clear from the context. For example, $\hat{F}(z^k)$ is associated with the random sample $\xi^k \in \Xi$. In addition, we denote $P_{\mathcal{Z}}(\cdot)$ as the *projection operator* onto the feasible set $\mathcal{Z}$.

4.2.1 The stochastic extra-point scheme

We first present the iterative updating rule for the stochastic extra-point scheme:

$$(4.2.3)\begin{cases} z^{k+0.5} := P_{\mathcal{Z}}\left(z^k + \beta(z^k - z^{k-1}) - \eta\hat{F}(z^k)\right), \\ z^{k+1} := P_{\mathcal{Z}}\left(z^k - \alpha\hat{F}(z^{k+0.5}) + \gamma(z^k - z^{k-1}) - \tau(\hat{F}(z^k) - \hat{F}(z^{k-1}))\right), \end{cases}$$

for $k = 0, 1, ...,$ where the sequence $\{z^k \mid k = 0, 1, ...\}$ is the sequence of iterates, and $\{z^{k+0.5} \mid k = 0, 1, ...\}$ is the sequence of *extra points*, which helps to produce the sequence of iterates.

In the case of *deterministic* strongly monotone VI, we introduced in Chapter 2 a *unifying* extra-point updating scheme, which includes specific first-order search directions such as the extra-gradient, the heavy-ball method, the optimistic method, and Nesterov's extrapolation; these are incorporated with the parameters $\alpha, \beta, \gamma, \eta, \tau \geq 0$. As any specific configuration of these parameters should be tailored to the problem structure at hand, our goal is to provide conditions of the parameters under which an *optimal* iteration complexity can be guaranteed. This line of analysis will now be extended to solve stochastic VI as given in (4.2.3). We shall first establish the relational inequalities between subsequent iterates in terms of the *expected* distance to the unique solution z^*, denote by $d_k = \mathbb{E}\left[\|z^k - z^*\|^2\right]$.

Lemma 4.2.1. *For the sequences* $\{z^k \mid k = 0, 1, ...\}$ *and* $\{z^{k+0.5} \mid k = 0, 1, ...\}$ *generated from the stochastic extra-point scheme* (4.2.3)*, the following inequality holds:*

$$\begin{aligned} & (1 - 4|\gamma - \beta| - \tau L) d_{k+1} \\ \leq \; & \left(1 + 4\gamma + 6|\gamma - \beta| + 4\tau L - \frac{1}{2}\alpha\mu\right) d_k + (2|\gamma - \beta| + 2\gamma + 4\tau L)\, d_{k-1} \\ & + \left(2\alpha^2 L^2 + 2|\gamma - \beta| + 2\gamma + \frac{5}{2}\alpha\mu - 1\right) \mathbb{E}\left[\|z^{k+0.5} - z^k\|^2\right] \\ & + 8\left(\alpha^2 + \frac{\tau}{L}\right)\sigma^2 + \frac{4\alpha\delta^2}{\mu} + 2(\eta - \alpha)\mathbb{E}\left[(z^{k+1} - z^{k+0.5})^\top \hat{F}(z^k)\right]. \end{aligned} \tag{4.2.4}$$

Proof. See Appendix 4.6.1. □

Lemma 4.2.1 forms a basis to the desired linear convergence, and it is possible to identify the conditions for the parameters $\alpha, \beta, \gamma, \eta, \tau$ in order to achieve linear convergence. Consider parameters satisfying

$$\eta = \alpha, \quad 2\alpha^2 L^2 + 2|\gamma - \beta| + 2\gamma + \frac{5}{2}\alpha\mu - 1 \leq 0, \tag{4.2.5}$$

and denote

$$\begin{cases} t_1 = \frac{1}{2}\alpha\mu - 4\gamma - 6|\gamma - \beta| - 4\tau L, \\ t_2 = 2|\gamma - \beta| + 2\gamma + 4\tau L, \quad t_3 = 4|\gamma - \beta| + \tau L. \end{cases} \tag{4.2.6}$$

Then we obtain from (4.2.4) that

$$(1 - t_3) d_{k+1} \;\leq\; (1 - t_1)\, d_k + t_2 d_{k-1} + 8\left(\alpha^2 + \frac{\tau}{L}\right)\sigma^2 + \frac{4\alpha\delta^2}{\mu}. \tag{4.2.7}$$

With additional constraints on t_1, t_2, t_3, the *variance-reduced* convergence result is summarized in the next theorem.

Theorem 4.2.2. *For non-negative parameters* $\alpha, \beta, \gamma, \eta, \tau$ *satisfying* (4.2.5) *and* (4.2.6)*, suppose that*

$$0 \leq t_3 < t_1 < 1, \quad t_2 < t_1 - t_3. \tag{4.2.8}$$

Let $q = \frac{2(1-t_3)}{t_1 - t_2 - t_3} > 1$*. For a fixed precision* $\epsilon > 0$*, denote* $K = \mathcal{O}\left(q \cdot \ln\left(\frac{1}{\epsilon}\right)\right)$*. Then we have*

$$d_K = \mathbb{E}\left[\|z^K - z^*\|^2\right] \leq \mathcal{O}(\epsilon) + \mathcal{O}(\sigma^2) + \mathcal{O}(\delta^2).$$

Proof. See Appendix 4.6.2. □

Regarding Theorem 4.2.2, we remark that a conventional way to achieve variance reduction is through increasing the mini-batch sample sizes. In fact, we may increase the sampel size linearly at a rate $\left(1+\frac{1}{q}\right)$ as k increases. We shall discuss more on this strategy in Section 4.3.

Next proposition concludes this subsection with a specific choice of the parameters.

Proposition 4.2.3. *If one chooses the following parameters*

$$(\alpha,\beta,\gamma,\eta,\tau)=\left(\frac{1}{4L},\frac{1}{128\kappa},\frac{1}{128\kappa},\frac{1}{4L},\frac{1}{128L\kappa}\right)$$

in (4.2.3) *(thus* $(t_1,t_2,t_3)=\left(\frac{1}{16\kappa},\frac{3}{64\kappa},\frac{1}{128\kappa}\right)$*) then it holds that*

$$d_k \le \left(1-\frac{1}{256\kappa}\right)^k \cdot \frac{269}{256}\|z^0-z^*\|^2+\mathcal{O}\left(\frac{\sigma^2}{L\mu}\right)+\mathcal{O}\left(\frac{\delta^2}{\mu^2}\right). \tag{4.2.9}$$

Proof. See Appendix 4.6.3. □

4.2.2 The stochastic extra-momentum scheme

In this subsection, we present an alternative stochastic first-order method that achieves the optimal iteration complexity as well, the *stochastic extra-momentum scheme*:

$$z^{k+1} := P_{\mathcal{Z}}\left(z^k-\alpha\hat{F}(z^k)+\gamma(z^k-z^{k-1})-\tau\left(\hat{F}(z^k)-\hat{F}(z^{k-1})\right)\right), \tag{4.2.10}$$

for $k=0,1,\ldots$.

Compared with the stochastic extra-point scheme (4.2.3), the above update (4.2.10) manipulates only the *momentum* terms alongside the stochastic gradient direction (the notion "gradient" here refers to the mapping $F(\cdot)$ in the VI model), namely the heavy-ball direction z^k-z^{k-1} and the optimism direction $\hat{F}(z^k)-\hat{F}(z^{k-1})$. Since it maintains a single sequence $\{z^k\}$ throughout the iterations, this scheme requires *one* projection per iteration, as compared to two projections in the case of the stochastic extra-point scheme. We shall remark that the method proposed in Kotsalis *et al.* [43] only considers the optimism term. Therefore, the stochastic extra-momentum scheme introduced above may be viewed as a generalization.

As in the previous subsection, we shall first establish a relational inequality between the iterates. As we can see from the lemma below, the structure of this relational inequality is in fact quite different from the previous case. The detailed proof can be found in the appendix.

Lemma 4.2.4. *For the sequence $\{z^k \mid k = 0, 1, ...\}$ generated from the stochastic extra-momentum scheme (4.2.10), the following inequality holds:*

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{2}+\frac{\alpha\mu}{2}-\frac{\gamma}{2}\right)\|z^{k+1}-z^*\|^2+\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k+1})\right)+\frac{1}{4}\|z^{k+1}-z^k\|^2\right]\\
\leq\ &\mathbb{E}\left[\frac{1}{2}\|z^k-z^*\|^2+\tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^k-z^{k-1}\|^2\right]\\
&+8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu}.
\end{aligned}$$

(4.2.11)

Proof. See Appendix 4.6.4. □

Observe that each of the terms on the LHS of (4.2.11) differs in the iteration index from the RHS exactly by one. This property enables us to design a possible *potential function* that measures the convergence of the iterative process. We shall specify additional conditions on the non-negative parameters α, γ, τ in order to further simplify (4.2.11):

$$1+\alpha\mu-\gamma \geq 1+\frac{\theta}{\kappa}, \quad \frac{\alpha}{\tau}=1+\frac{\theta}{\kappa}, \quad \frac{1}{8\tau^2L^2+2\gamma} \geq 1+\frac{\theta}{\kappa}, \tag{4.2.12}$$

where $\theta \in (0, 1]$ is some constant independent of κ. Note that the LHS of each inequality in (4.2.12) is the ratio between the coefficients on the LHS and RHS of (4.2.11) for each corresponding term. Therefore, the relation (4.2.11) can be rearranged as:

$$\begin{aligned}
&\left(1+\frac{\theta}{\kappa}\right)\mathbb{E}\left[\frac{1}{2}\|z^{k+1}-z^*\|^2+\tau(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k+1})\right)\right.\\
&\left.+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^{k+1}-z^k\|^2\right]\\
\leq\ &\mathbb{E}\left[\left(\frac{1}{2}+\frac{\alpha\mu}{2}-\frac{\gamma}{2}\right)\|z^{k+1}-z^*\|^2+\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k+1})\right)+\frac{1}{4}\|z^{k+1}-z^k\|^2\right]\\
\leq\ &\mathbb{E}\left[\frac{1}{2}\|z^k-z^*\|^2+\tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^k-z^{k-1}\|^2\right]\\
&+8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu}.
\end{aligned}$$

(4.2.13)

Now, by defining the potential function V_k as

$$V_k=\mathbb{E}\left[\frac{1}{2}\|z^k-z^*\|^2+\tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^k-z^{k-1}\|^2\right],$$

inequality (4.2.13) can be rewritten as

$$\left(1+\frac{\theta}{\kappa}\right)V_{k+1} \leq V_k+8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu}. \tag{4.2.14}$$

This leads to our final results, as summarized in the next theorem:

Theorem 4.2.5. *Suppose that the non-negative parameters α, γ, τ satisfy (4.2.12) for some constant $\theta \in (0, 1]$. For the sequence $\{z^k \mid k = 0, 1, ...\}$ generated from the stochastic extra-momentum scheme (4.2.10), the expected distance to the solution of iterate z^k is bounded by:*

$$\mathbb{E}\left[\|z^k - z^*\|^2\right] \leq 2\left(1 + \frac{\theta}{\kappa}\right)^{-k} \|z^0 - z^*\|^2 + \left(\frac{\kappa}{\theta} + 1\right) \cdot 32\tau^2\sigma^2 + \frac{2\kappa\alpha\delta^2}{\theta\mu}. \tag{4.2.15}$$

Proof. See Appendix 4.6.5. □

A simple choice of parameters leads to:

Proposition 4.2.6. *If we choose the parameters as*

$$(\alpha, \tau, \gamma) = \left(\frac{1}{4L}, \frac{\alpha}{1 + \frac{\theta}{\kappa}}, \frac{1}{8(\kappa + \theta)}\right), \quad \theta = \frac{1}{8},$$

then scheme (4.2.10) assures that

$$d_k \leq 2\left(1 - \frac{1}{8\kappa + 1}\right)^k \|z^0 - z^*\|^2 + \frac{128\sigma^2}{\mu(8L + \mu)} + \frac{4\delta^2}{\mu^2}. \tag{4.2.16}$$

Proof. It follows by substituting the parameter choice into (4.2.15). □

This shows that if we run the stochastic extra-momentum scheme (4.2.10) with the above parameter choice, then in $K = \mathcal{O}(\kappa \ln \frac{1}{\epsilon})$ iterations we will reach a solution satisfying

$$\mathbb{E}\left[\|z^K - z^*\|^2\right] \leq \mathcal{O}(\epsilon) + \mathcal{O}\left(\frac{\sigma^2}{L\mu}\right) + \mathcal{O}\left(\frac{\delta^2}{\mu^2}\right). \tag{4.2.17}$$

In view of (4.2.9) and (4.2.17), both stochastic extra-point and stochastic extra-momentum schemes guarantee

$$d_K = \mathbb{E}\left[\|z^K - z^*\|^2\right] \leq \mathcal{O}\left(\epsilon + \frac{\sigma^2}{L\mu} + \frac{\delta^2}{\mu^2}\right) \tag{4.2.18}$$

after $K = \mathcal{O}\left(\kappa \ln \frac{1}{\epsilon}\right)$ iterations. Note that in the analysis of this section we focus on reducing the deterministic error $\|z^0 - z^*\|^2$ by applying fixed parameters, while *assuming the variance is controllable* through some variance reduction technique (see the discussion in Section 4.3 for example). However, if we aim to reduce the stochastic error σ^2 throughout the iterations alongside the deterministic error without increasing samples per iteration, then decreasing parameters should be considered. In particular, choosing

$$(\alpha_k, \beta_k, \gamma_k, \eta_k, \tau_k) = \left(\frac{4}{(k + 16\kappa + 2)\mu}, \frac{\alpha_k^2\mu^2}{512}, \frac{\alpha_k^2\mu^2}{512}, \frac{4}{(k + 16\kappa + 2)\mu}, \frac{\alpha_k^2\mu}{512\kappa}\right),$$

for the stochastic extra-point scheme, and

$$(\alpha_k, \tau_k, \gamma_k) = \left(\frac{3}{\mu(k+9\kappa)}, \frac{3}{\mu(k+9\kappa+1)}, \frac{1}{k+9\kappa+1} \right)$$

for the stochastic extra-momentum scheme, a *uniform sublinear convergence* can be established for both schemes in the form:

$$d_k \leq \mathcal{O}\left(\frac{Q}{k+\kappa} + \frac{\delta^2}{\mu^2} \right), \quad Q = \mathcal{O}\left(\max\left\{ \kappa \|z^0 - z^*\|^2, \frac{\sigma^2}{\mu^2} \right\} \right). \tag{4.2.19}$$

Appendix 4.7 provides detailed derivations for the stochastic extra-point scheme, and the proofs for the stochastic extra-momentum scheme can be derived via a similar logic. Let us consider combining both strategies: running $K_1 = \mathcal{O}\left(\kappa \ln \frac{1}{\epsilon}\right)$ iterations with fixed parameters followed by $K_2 = \mathcal{O}\left(\frac{\sigma^2}{\epsilon\mu^2}\right)$ iterations with decreasing parameters. Then we have:

$$d_{K_1+K_2} \leq \mathcal{O}\left(\frac{Q}{K_2+\kappa} + \frac{\delta^2}{\mu^2} \right),$$

where

$$Q = \mathcal{O}\left(\max\left\{ \kappa \|z^{K_1} - z^*\|^2, \frac{\sigma^2}{\mu^2} \right\} \right) = \mathcal{O}\left(\kappa\epsilon + \frac{\kappa\delta^2}{\mu^2} + \frac{\sigma^2}{\mu^2} \right).$$

It is then clear that $d_{K_1+K_2} \leq \mathcal{O}\left(\epsilon + \frac{\delta^2}{\mu^2}\right)$, with total iterations $K_1 + K_2$ $= \mathcal{O}\left(\kappa \ln \frac{1}{\epsilon} + \frac{\sigma^2}{\epsilon\mu^2}\right) = \mathcal{O}\left(\max\left\{\kappa \ln \frac{1}{\epsilon}, \frac{\sigma^2}{\epsilon\mu^2}\right\}\right)$, which matches the optimal iteration complexity established in [43] for a simultaneous reduction of both deterministic and stochastic errors.

4.3 A Stochastic Zeroth-Order Approach to Saddle-Point Problems

In this section, we shall apply the proposed stochastic extra-point/extra-momentum scheme to solve the following saddle-point problem without needing to compute the gradients of f:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \quad f(x, y), \tag{4.3.20}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^m$ are convex sets, $f(x, y)$ is strongly convex (with fixed y) and strongly concave (with fixed x) with modulus μ, and the partial gradients $\nabla_x f(x, y)/\nabla_y f(x, y)$ are Lipschitz continuous with constant L_x/L_y for fixed y/x, and with constant L_{xy} with fixed x/y. We let $L = 2 \cdot \max(L_x, L_y, L_{xy})$. We further assume that the function $f(x, y)$ is Lipschitz continuous for either fixed x or y with constant M. This

implies that the norms of the partial gradients are bounded by M: $\|\nabla_x f(x,y)\| \leq M$, $\|\nabla_y f(x,y)\| \leq M$. In particular, we consider the settings when the partial gradients $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ (and any higher-order information) are not available. Furthermore, the *exact* evaluation of the function value itself is also not available; instead, we can only access a *stochastic oracle* $\hat{f}(x,y,\xi)$, which satisfies the following assumption:

$$
\begin{cases}
\mathbb{E}\left[\hat{f}(x,y,\xi)\right] = f(x,y), \\
\mathbb{E}\left[\nabla_x \hat{f}(x,y,\xi)\right] = \nabla_x f(x,y), \quad \mathbb{E}\left[\nabla_y \hat{f}(x,y,\xi)\right] = \nabla_y f(x,y), \\
\mathbb{E}\left[\|\nabla_x \hat{f}(x,y,\xi) - \nabla_x f(x,y)\|^2\right] \leq \sigma^2, \\
\mathbb{E}\left[\|\nabla_y \hat{f}(x,y,\xi) - \nabla_y f(x,y)\|^2\right] \leq \sigma^2.
\end{cases}
\tag{4.3.21}
$$

Now, we shall use the so-called *smoothing* technique to approximate the first-order information, which then enables us to apply the proposed stochastic methods for VI, which includes the saddle-point model as a special case. In particular, we use a randomized smoothing scheme using uniform distributions U_b/V_b over the unit Euclidean ball B in the $\mathbb{R}^n/\mathbb{R}^m$ space, respectively. The smoothing functions with parameters $\rho_x, \rho_y > 0$ are defined as follows:

$$
\begin{aligned}
f_{\rho_x}(x,y) &:= \mathbb{E}_{u\sim U_b}\left[f(x+\rho_x u, y)\right] = \frac{1}{\alpha(n)}\int_B f(x+\rho_x u, y)du, \\
f_{\rho_y}(x,y) &:= \mathbb{E}_{v\sim V_b}\left[f(x, y+\rho_y v)\right] = \frac{1}{\alpha(m)}\int_B f(x, y+\rho_y v)dv,
\end{aligned}
$$

where $\alpha(n)/\alpha(m)$ is the volume of the unit ball in $\mathbb{R}^n/\mathbb{R}^m$.

Let us summarize main properties of the smoothing functions f_{ρ_x}, f_{ρ_y} below:

Lemma 4.3.1. *Let U_{S_p}/V_{S_p} be the uniform distribution on the unit sphere S_p in $\mathbb{R}^n/\mathbb{R}^m$. Then the smoothing functions f_{ρ_x}, f_{ρ_y} are continuously differentiable, and their partial gradients $\nabla_x f_{\rho_x}, \nabla_y f_{\rho_y}$ can be expressed as:*

$$
\begin{aligned}
\nabla_x f_{\rho_x}(x,y) &:= \mathbb{E}_{u\sim U_{S_p}}\left[\frac{n}{\rho_x} f(x+\rho_x u, y)u\right] = \frac{1}{\beta(n)}\int_{u\in S_p}\frac{n}{\rho_x}\left(f(x+\rho_x u, y) - f(x,y)\right)udu, \\
\nabla_y f_{\rho_y}(x,y) &:= \mathbb{E}_{v\sim V_{S_p}}\left[\frac{m}{\rho_y} f(x, y+\rho_y v)v\right] = \frac{1}{\beta(m)}\int_{v\in S_p}\frac{m}{\rho_y}\left(f(x, y+\rho_y v) - f(x,y)\right)vdv,
\end{aligned}
$$

where $\beta(n)/\beta(m)$ is the surface area of the unit sphere in $\mathbb{R}^n/\mathbb{R}^m$.

Furthermore, for any $x \in \mathbb{R}^n$ and any $y \in \mathbb{R}^m$, we have:

$$
\|\nabla_x f_{\rho_x}(x,y) - \nabla_x f(x,y)\| \leq \frac{\rho_x n L}{2}, \quad \|\nabla_y f_{\rho_y}(x,y) - \nabla_y f(x,y)\| \leq \frac{\rho_y m L}{2},
\tag{4.3.22}
$$

(4.3.23)

$$\begin{cases} \mathbb{E}_{u\sim U_{S_p}}\left[\left\|\frac{n}{\rho_x}\left(f(x+\rho_x u,y)-f(x,y)\right)u\right\|^2\right] \le 2n\|\nabla_x f(x,y)\|^2+\frac{\rho_x^2L^2n^2}{2}, \\ \\ \mathbb{E}_{v\sim V_{S_p}}\left[\left\|\frac{m}{\rho_y}\left(f(x,y+\rho_y v)-f(x,y)\right)v\right\|^2\right] \le 2m\|\nabla_y f(x,y)\|^2+\frac{\rho_y^2L^2m^2}{2}. \end{cases}$$

Proof. For the first half of the statement, cf. [87] (Lemma 4.4), and for the second half, cf. [20] (proofs for Propositions 2.7.5 and 2.7.6). Note that the proofs for the minimax function follows simply by fixing one of the two variables. □

We are now ready to define the *stochastic zeroth-order gradient* as follows:

$$\begin{cases} F_{\rho_x}(x,y,\xi,u) := \frac{n}{\rho_x}\left(\hat{f}(x+\rho_x u,y,\xi)-\hat{f}(x,y,\xi)\right)u, \\ F_{\rho_y}(x,y,\xi,v) := \frac{m}{\rho_y}\left(\hat{f}(x,y+\rho_y v,\xi)-\hat{f}(x,y,\xi)\right)v, \end{cases} \tag{4.3.24}$$

where u and v are the uniformly distributed random vectors over the unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively.

The next lemma shows that such stochastic zeroth-order gradients are unbiased with respect to the gradients of the smoothing functions and have uniformly bounded variance.

Lemma 4.3.2. *The stochastic zeroth-order gradients defined in* (4.3.24) *are unbiased and have bounded variance for all* $(x,y)\in\mathcal{X}\times\mathcal{Y}$*:*

$$\mathbb{E}_{\xi,u}\left[F_{\rho_x}(x,y,\xi,u)\right]=\nabla_x f_{\rho_x}(x,y), \quad \mathbb{E}_{\xi,v}\left[F_{\rho_y}(x,y,\xi,v)\right]=\nabla_y f_{\rho_y}(x,y), \tag{4.3.25}$$

and

$$\begin{cases} \mathbb{E}_{\xi,u}\left[\|F_{\rho_x}(x,y,\xi,u)-\nabla_x f_{\rho_x}(x,y)\|^2\right]\le\tilde{\sigma}^2, \\ \\ \mathbb{E}_{\xi,v}\left[\|F_{\rho_y}(x,y,\xi,v)-\nabla_y f_{\rho_y}(x,y)\|^2\right]\le\tilde{\sigma}^2, \end{cases} \tag{4.3.26}$$

where $\tilde{\sigma}^2=2\cdot\max\left\{nM^2+n\sigma^2+n^2\rho_x^2L^2,\, mM^2+m\sigma^2+m^2\rho_y^2L^2\right\}$.

Proof. See Appendix 4.6.6. □

Before applying the stochastic extra-point/extra-momentum scheme to solve (4.3.20), let us first introduce the connections between these two models. As we regard the saddle-point model as a special case of VI, we shall treat the variables x, y in the saddle-point problem

as one variable and denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, z = (x, y)$. Additionally, we define:

$$F(z) := \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}, \quad F_\rho(z) := \begin{pmatrix} \nabla_x f_{\rho_x}(x,y) \\ -\nabla_y f_{\rho_y}(x,y) \end{pmatrix},$$
$$\hat{F}_\rho(z,u,v,\xi) := \begin{pmatrix} F_{\rho_x}(x,y,\xi,u) \\ -F_{\rho_y}(x,y,\xi,v) \end{pmatrix}.$$

These terms correspond to the gradient of $f(x, y)$, the gradient of the smoothing functions $f_{\rho_x}(x, y)$ and $f_{\rho_y}(x, y)$, and the stochastic zeroth-order gradient, respectively. Note that we have flipped the sign on partial gradient correspond to y to account for the *concavity* of f with respect to y.

Finally, as we shall use a sample size of $t_k \in \mathbb{N}$ (a natural number) at iteration k, we reserve the subscripts for the random vectors ξ, u, v for the sample index (i), $i = 1, ..., t_k$, and denote:

$$\hat{F}_\rho^k(z^k) = \frac{1}{t_k} \sum_{i=1}^{t_k} \hat{F}_\rho(z^k, u_{(i)}^k, v_{(i)}^k, \xi_{(i)}^k).$$

In the above definition we suppress the notation of the random vectors u, v, ξ on the LHS for cleaner presentation. Note that by the *law of large numbers*, together with (4.3.25)-(4.3.26), we have

$$\mathbb{E}\left[\hat{F}_\rho^k(z^k)\right] = F_\rho(z^k), \quad \mathbb{E}\left[\|\hat{F}_\rho^k(z^k) - F_\rho(z^k)\|^2\right] \leq \frac{2\tilde{\sigma}^2}{t_k}. \tag{4.3.27}$$

4.3.1 Sample complexity analysis: stochastic zeroth-order extra-point method

Recall our objective in (4.3.20). With only noisy function value $\hat{f}(x, y, \xi)$ accessible, we propose the *stochastic zeroth-order extra-point method* that updates $(x, y) := z$ simultaneously with the following update rule:

$$\begin{cases} z^{k+0.5} & := & P_{\mathcal{Z}}\left(z^k + \beta(z^k - z^{k-1}) - \eta \hat{F}_\rho^k(z^k)\right), \\ z^{k+1} & := & P_{\mathcal{Z}}\left(z^k - \alpha \hat{F}_\rho^{k+0.5}(z^{k+0.5}) + \gamma(z^k - z^{k-1}) - \tau\left(\hat{F}_\rho^k(z^k) - \hat{F}_\rho^{k-1}(z^{k-1})\right)\right). \end{cases} \tag{4.3.28}$$

Compare the above update with its original variant in (4.2.3) for solving stochastic VI, the update direction $\hat{F}(z^k)$ is replaced by the averaged stochastic zeroth-order gradient $\hat{F}_\rho^k(z^k)$ with sample size t_k (similarly $\hat{F}(z^{k+0.5})$ is replaced by $\hat{F}_\rho^{k+0.5}(z^{k+0.5})$ with sample size $t_{k+0.5}$). This circumvents the inaccessible first-order information and equips us with appropriate tools to reduce the variance and achieve overall linear convergence.

The next lemma establishes the relational inequality between the subsequent iterates in terms of the expected distance to the solution $d_k = \mathbb{E}\left[\|z^k - z^*\|^2\right]$, similar to what we did in Section 4.2.1. The differences lie in the corresponding stochastic error terms shown below. Note that in each iteration we take two batches of samples t_k and $t_{k+0.5}$. The batch size t_{k-1} also appears because the iterate z^{k-1} is used in each iteration.

Lemma 4.3.3. *For the sequences $\{z^k \mid k = 0, 1, ...\}$ and $\{z^{k+0.5} \mid k = 0, 1, ...\}$ generated from the stochastic zeroth-order extra-point method (4.3.28), the following inequality holds:*

$$\begin{aligned}
& (1 - 4|\gamma - \beta| - \tau L) d_{k+1} \\
\leq\ & \left(1 + 4\gamma + 6|\gamma - \beta| + 4\tau L - \frac{1}{2}\alpha\mu\right) d_k + (2|\gamma - \beta| + 2\gamma + 4\tau L)\, d_{k-1} \\
& + \left(\alpha^2 L^2 + 2|\gamma - \beta| + 2\gamma + \frac{5}{2}\alpha\mu - 1\right) \mathbb{E}\left[\|z^{k+0.5} - z^k\|^2\right] \\
& + 16\tilde{\sigma}^2 \left(\left(\alpha^2 + \frac{\tau}{L}\right)\frac{1}{t_k} + \frac{\alpha^2}{t_{k+0.5}} + \frac{\tau}{L t_{k-1}}\right) + 4\left(\tau L + \alpha^2 L^2 + \frac{\alpha L^2}{4\mu}\right)(\rho_x^2 n^2 + \rho_y^2 m^2) \\
& + 2(\eta - \alpha)\mathbb{E}\left[(z^{k+1} - z^{k+0.5})^\top \hat{F}_\rho^k(z^k)\right]. \qquad (4.3.29)
\end{aligned}$$

Proof. See Appendix 4.6.7. □

With the relational inequality in Lemma 4.3.3, we shall adopt the *same* conditions: (4.2.5), (4.2.6), and (4.2.8) for the parameters $\alpha, \beta, \gamma, \eta, \tau$. Therefore, the results in Theorem 4.2.2 are directly applicable. In addition, we are now equipped with the variable sample size $t_k/t_{k+0.5}$ to control the variance terms, as well as the smoothing parameters ρ_x, ρ_y to control the bias terms.

We shall utilize the example in Proposition 4.2.3 to analyze the sample complexity of the proposed method. The result is provided in the next proposition:

Proposition 4.3.4 (Sample complexity result 1)**.** *The stochastic zeroth-order extra-point method (4.3.28) with the following parameter choice:*

$$(\alpha, \beta, \gamma, \eta, \tau) = \left(\frac{1}{4L}, \frac{1}{128\kappa}, \frac{1}{128\kappa}, \frac{1}{4L}, \frac{1}{128L\kappa}\right)$$

and

$$\begin{aligned}
& t_{k+0.5} = t_k = \left\lceil \left(1 - \frac{1}{256\kappa}\right)^{-k} \right\rceil, \quad \rho_x = \frac{1}{\sqrt{2}n\kappa}\left(1 - \frac{1}{256\kappa}\right)^{\frac{K}{2}}, \\
& \rho_y = \frac{1}{\sqrt{2}m\kappa}\left(1 - \frac{1}{256\kappa}\right)^{\frac{K}{2}}, \qquad (4.3.30)
\end{aligned}$$

where K is the iteration count decided in advance, outputs z^K such that $d_K = \mathbb{E}\left[\|z^K - z^*\|^2\right] \leq \epsilon$, with $K = \mathcal{O}\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$, and the total sample complexity of the procedure

is

$$\sum_{k=0}^{K-1}(t_k + t_{k+0.5}) = \mathcal{O}\left(\frac{\kappa\|z^0 - z^*\|^2}{\epsilon} + \frac{\tilde{\sigma}^2}{\epsilon\mu^2}\right).$$

Proof. See Appendix 4.6.8. □

4.3.2 Sample complexity: stochastic zeroth-order extra-momentum method

Next we consider the *stochastic zeroth-order extra-momentum method*, with one projection per each iteration:

$$z^{k+1} := P_{\mathcal{Z}}\left(z^k - \alpha\hat{F}_\rho^k(z^k) + \gamma(z^k - z^{k-1}) - \tau\left(\hat{F}_\rho^k(z^k) - \hat{F}_\rho^{k-1}(z^{k-1})\right)\right). \tag{4.3.31}$$

The relational inequality, similar to Lemma 4.2.4, is established in the next lemma:

Lemma 4.3.5. *For the sequence $\{z^k \mid k = 0, 1, ...\}$ generated from the stochastic zeroth-order extra-momentum method (4.3.31), the following inequality holds:*

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{2} + \frac{\alpha\mu}{2} - \frac{\gamma}{2}\right)\|z^{k+1} - z^*\|^2 + \alpha(z^{k+1} - z^*)^\top\left(\hat{F}_\rho^k(z^k) - \hat{F}_\rho^{k+1}(z^{k+1})\right)\right.\\
&\left.+\frac{1}{4}\|z^{k+1} - z^k\|^2\right]\\
\leq\ &\mathbb{E}\left[\frac{1}{2}\|z^k - z^*\|^2 + \tau(z^k - z^*)^\top\left(\hat{F}_\rho^{k-1}(z^{k-1}) - \hat{F}_\rho^k(z^k)\right) + \left(2\tau^2L^2 + \frac{\gamma}{2}\right)\|z^k - z^{k-1}\|^2\right]\\
&+16\tau^2\tilde{\sigma}^2\left(\frac{1}{t_k} + \frac{1}{t_{k-1}}\right) + L^2\left(4\tau^2 + \frac{\alpha}{8\mu}\right)(\rho_x^2n^2 + \rho_y^2m^2).
\end{aligned} \tag{4.3.32}$$

Proof. See Appendix 4.6.9. □

With the *same* condition as in (4.2.12) for the parameters α, γ, τ, we can derive the similar bound to (4.2.13) (with $\hat{F}(z^k)$ replaced with $\hat{F}_\rho^k(z^k)$ and with the new stochastic error expression) and define the potential function:

$$V_k = \mathbb{E}\left[\frac{1}{2}\|z^k - z^*\|^2 + \tau(z^k - z^*)^\top\left(\hat{F}_\rho^{k-1}(z^{k-1}) - \hat{F}_\rho^k(z^k)\right) + \left(2\tau^2L^2 + \frac{\gamma}{2}\right)\|z^k - z^{k-1}\|^2\right].$$

Therefore, the following inequality holds:

$$\left(1 + \frac{\theta}{\kappa}\right)V_{k+1} \leq V_k + 16\tau^2\tilde{\sigma}^2\left(\frac{1}{t_k} + \frac{1}{t_{k-1}}\right) + L^2\left(4\tau^2 + \frac{\alpha}{8\mu}\right)(\rho_x^2n^2 + \rho_y^2m^2), \tag{4.3.33}$$

and we can apply the results directly from Theorem 4.2.5. In addition, with increasing sample sizes t_k and the smoothing parameters ρ_x, ρ_y, we are able to control the bias and the variance terms in the above inequality. We give the results of sample complexity in the next proposition.

Proposition 4.3.6 (Sample complexity result 2)**.** *The stochastic zeroth-order extra-momentum method* (4.3.31) *with the following parameter choice:*

$$(\alpha, \tau, \gamma) = \left(\frac{1}{4L}, \frac{\alpha}{1 + \frac{\theta}{\kappa}}, \frac{1}{8(\kappa + \theta)} \right), \quad \theta = \frac{1}{8},$$

and

$$t_k = \left\lceil \left(1 - \frac{1}{8\kappa + 1}\right)^{-k} \right\rceil, \quad \rho_x = \frac{1}{\sqrt{2}n\kappa} \left(1 - \frac{1}{8\kappa + 1}\right)^{\frac{K}{2}},$$
$$\rho_y = \frac{1}{\sqrt{2}m\kappa} \left(1 - \frac{1}{8\kappa + 1}\right)^{\frac{K}{2}},$$

where K is the iteration count decided in advance, outputs z^K such that $d_K = \mathbb{E}\left[\|z^K - z^*\|^2\right] \leq \epsilon$, with $K = \mathcal{O}\left(\kappa \ln\left(\frac{1}{\epsilon}\right)\right)$ and the total sample complexity of the procedure is

$$\sum_{k=0}^{K-1} t_k = \mathcal{O}\left(\frac{\kappa \|z^0 - z^*\|^2}{\epsilon} + \frac{\tilde{\sigma}^2}{\epsilon \mu^2} \right).$$

Proof. See Appendix 4.6.10. □

To make the sample complexities given by Proposition 4.3.4 and 4.3.6 more explicit, let us apply the specific choices of ρ_x, ρ_y in these propositions and the expression of $\tilde{\sigma}^2$ given in Lemma 4.3.2, which results in the following bound:

$$\mathcal{O}\left(\frac{\kappa \|z^0 - z^*\|^2}{\epsilon} + \frac{(n+m)M^2}{\epsilon \mu^2} \right). \tag{4.3.34}$$

Note that we took $\sigma^2 = 0$ in the above derivation (4.3.34) (i.e. the function value and the gradients can be evaluated exactly). We also remind that M is the Lipschitz constant of the function $f(x, y)$ and $\kappa = \frac{L}{\mu}$. Therefore, (4.3.34) gives an upper bound on the sample complexity for the black-box strongly-convex-strongly-concave saddle-point problem, obtained through the proposed stochastic extra-point/extra-momentum methods. While the tight sample complexity bound for black-box convex optimization has been established (cf. [17, 88]), a good or even tight lower bound on the sample complexity for saddle-point problems is unavailable at this point. Therefore, the result in (4.3.34) serves as a reasonable upper bound on the sample complexity for the research to follow along this line.

4.4 Numerical Experiments

In this section, we present the results of three sets of experiments to demonstrate the potential advantages of combined search directions as in the proposed stochastic extra-point and stochastic extra-momentum schemes. The first set of experiments relates to a regularized two-player zero-sum matrix game with an *uncertain* payoff matrix. In particular, the payoff matrix A_ξ is randomly distributed and can only be sampled for each (mixed) strategy. The problem can be formulated as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} \quad & f(x,y) = \mathbb{E}\left[\frac{\lambda_x}{2}\|x\|^2 + x^\top A_\xi y - \frac{\lambda_y}{2}\|y\|^2\right] \\ \text{s.t.} \quad & \sum_{i=1}^n x_i = 1, \quad \sum_{j=1}^m y_j = 1, \quad x, y \geq 0. \end{aligned} \tag{4.4.35}$$

We sample the matrix A_ξ element-wise i.i.d with $A_\xi \sim \mathcal{N}\left(A_0, \sigma^2 I_{(n+m)}\right)$ with randomly generated A_0. The problem parameters are set as: $n = 10$, $m = 20$, $\sigma^2 = 1$, $\lambda_x = \lambda_y = \lambda = 0.01$, $\kappa = \frac{L}{\mu} \approx 1.61\cdot 10^4$, where L and μ are the largest and the smallest singular values of the Jacobian matrix $\begin{pmatrix} \lambda I_n & A_0 \\ -A_0^\top & \lambda I_m \end{pmatrix}$ respectively. The duality gap at $\max_{y\in\mathcal{Y}} f(x^k, y) - \min_{x\in\mathcal{X}} f(x, y^k)$ is used as the merit function with $\mathcal{X}, \mathcal{Y}$ being the corresponding simplex constraints.

The second set of experiments is designed to test on a strongly-convex-strongly-concave stochastic saddle-point problem with more general polyhedral constraints:

$$\begin{aligned} \min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} \quad & f(x,y) = \mathbb{E}\left[\frac{\lambda_x}{2}\|x\|^2 + x^\top A_\xi y - \frac{\lambda_y}{2}\|y\|^2\right] \\ \text{s.t.} \quad & A_1 x \leq b_1, \quad C_1 y \leq d_1. \end{aligned} \tag{4.4.36}$$

The stochastic matrix A_ξ is sampled in the same way as in the first experiment. The constraint parameters $A_1 \in \mathbb{R}^{(2n+2)\times n}$, $b_1 \in \mathbb{R}^{2n+2}$, $C_1 \in \mathbb{R}^{(2m+2)\times m}$, $d_1 \in \mathbb{R}^{2m+2}$ are generated randomly such that the initial iterate $x_0 = e_1, y_0 = e_1$ is strictly feasible. We set the variance in this experiment as $\sigma^2 = 0.2$, while leaving other parameters the same as in the first experiment. The duality gap is used as the merit function.

The third set of experiments is designed to solve a stochastic linear complementarity problem (LCP):

$$z^* \geq 0, \quad F(z^*) \geq 0, \quad (z^*)^\top F(z^*) = 0, \tag{4.4.37}$$

where $F(z) = Mz + q$ is a randomly generated strongly-monotone operator and $M = Q + P$ is a summation of a positive diagonal matrix Q and a skew-symmetric matrix P, and we

add stochastic noises with element-wise i.i.d $\mathcal{N}\left(0, \sigma^2\right)$ to M. Other problem parameters are set as: $n = 20$ (variable size), $\kappa \approx 40.7$, $\sigma^2 = 1$.

In the above three experiments, we first observe the convergence for different methods in the deterministic ($\sigma^2 = 0$) case, with the parameters manually tuned to their best performance. We then add the stochastic noises to the problem and observe the average performance of ten runs, with parameters for each method fixed from the previous deterministic experiments. The results are shown in Figure 4.1 and Figure 4.2 respectively, and the parameters are recorded in Table 4.1 for references. In these problem instances, there are parameter combinations such that the more general schemes (extra-point, extra-momentum) can outperform specific components of their own (extra-gradient, OGDA, heavy-ball). This demonstrates that indeed *there exist problem structures where the numerically best performing method can only be found in the non-trivial configurations of some existing different methods but not any individual one of them.* We shall remark here that the point of such kind of experiments is to show that a combined scheme *can* be superior than each individual method alone. The experiments are however not meant to address the issue of *how* to find superior hybrids of the methods, which on its own is a future research topic. Furthermore, a potentially superior performance may come at the cost of some parameter search. Suitable setting of parameters may vary from problem classes to problem classes and should be itself considered as *a learning process.*

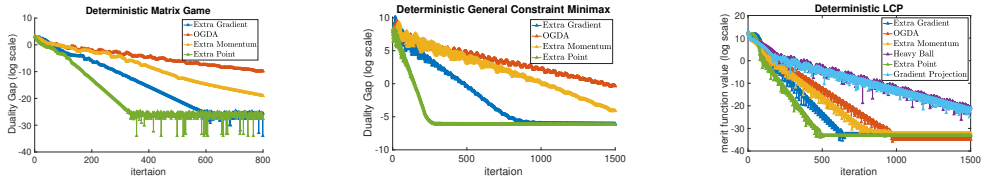

Figure 4.1: Convergence in deterministic problems

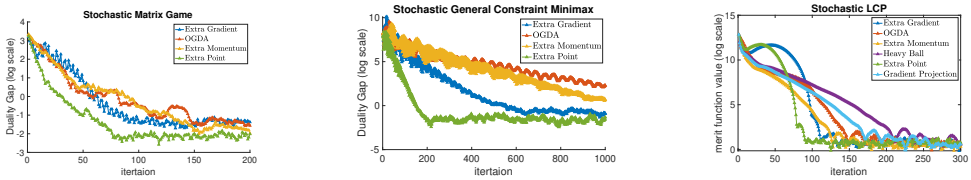

Figure 4.2: Convergence in stochastic problems

Problem	Matrix Game ($\sigma^2 = 1$)	General Constraint ($\sigma^2 = 0.2$)	LCP ($\sigma^2 = 1$)
Proj.	N/A	N/A	$0.0235/-/-/-/-$
HB	N/A	N/A	$0.0188/-/-/0.0146/-$
EG	$0.015/0.015/-/-/-$	$0.0061/0.0061/-/-/-$	$0.034/0.034/-/-/-$
OGDA	$0.007/-/-/-/0.007$	$0.0035/-/-/-/0.0035$	$0.024/-/-/-/0.0234$
ExM	$0.005/-/-/0.11/0.01$	$0.0023/-/-/0.16/0.0043$	$0.024/-/-/0.216/0.0234$
ExP	$0.01/0.024/0.004/0.004/0.001$	$0.0032/0.04/0.0176/0.22/0.0007$	$0.034/0.34/0.0323/0.34/0.0068$

Table 4.1: Parameter choices for different problems in the order of $\alpha/\eta/\beta/\gamma/\tau$. N/A: no obvious convergence. σ^2: variance of the stochastic noise. Proj.: gradient projection method; HB: heavy-ball; EG: extra-gradient; ExM: extra-momentum; ExP: extra-point.

4.5 Conclusion

In this chapter, we propose two new schemes of stochastic first-order methods to solve strongly monotone VI problems: the stochastic extra-point scheme and the stochastic extra-momentum scheme. The first scheme features a high flexibility in the configuration of parameter choices that can be tailored to different problem classes. The second scheme has the advantage of maintaining a single sequence throughout the iterations thus requiring only one projection per iteration, as opposed to most other first method that maintains an extra iterative sequence. Both methods achieve optimal iteration complexity bound, provided that the stochastic gradient oracle allows the variance to be controllable. The application of these two schemes to solve stochastic black-box saddle-point problem is also presented. Through a randomized smoothing scheme, the stochastic oracles required in these two schemes can be constructed via stochastic zeroth-order gradient approximation. The variance is thus controllable by mini-batch sampling with linearly increasing sample sizes per iteration, and the sample complexity results are derived. Finally, preliminary numerical experiments show that there exist different problem classes where combined search directions is preferable over any individual ones in our schemes.

4.6 Appendix A: Proofs of technical results

4.6.1 Proof of Lemma 4.2.1

First of all, by the 1-co-coerciveness (cf. e.g. Proposition 4.4 in [5]) of the projection operator $P_{\mathcal{Z}}$, we have

$$\begin{aligned}
& \|z^{k+1} - z^*\|^2 \\
\leq\ & (z^{k+1} - z^*)^\top \left(z^k - \alpha \hat{F}(z^{k+0.5}) + \gamma(z^k - z^{k-1}) - \tau \left(\hat{F}(z^k) - \hat{F}(z^{k-1}) \right) - z^* \right) \\
=\ & \frac{1}{2}\|z^{k+1} - z^*\|^2 + \frac{1}{2}\|z^k - z^*\|^2 - \frac{1}{2}\|z^{k+1} - z^k\|^2 - \tau(z^{k+1} - z^*)^\top \left(\hat{F}(z^k) - \hat{F}(z^{k-1}) \right) \\
& + (z^{k+1} - z^*)^\top \left(-\alpha \hat{F}(z^{k+0.5}) + \gamma(z^k - z^{k-1}) \right).
\end{aligned} \tag{4.6.38}$$

By adding and subtracting $z^{k+0.5}$ in $z^{k+1} - z^*$, we shall decompose the last term in the above inequality into $(a) + (b) + (c)$, where

$$\begin{aligned}
(a) &= (z^{k+1} - z^{k+0.5})^\top \left(-\eta \hat{F}(z^k) + \beta(z^k - z^{k-1}) \right) \\
(b) &= (z^{k+1} - z^{k+0.5})^\top \left(-\alpha \hat{F}(z^{k+0.5}) + \eta \hat{F}(z^k) + (\gamma - \beta)(z^k - z^{k-1}) \right) \\
(c) &= (z^{k+0.5} - z^*)^\top \left(-\alpha \hat{F}(z^{k+0.5}) + \gamma(z^k - z^{k-1}) \right).
\end{aligned} \tag{4.6.39}$$

Let us first use the optimality condition of $z^{k+0.5}$ to bound term (a):

$$\langle z^{k+0.5} - z^k - \beta(z^k - z^{k-1}) + \eta \hat{F}(z^k), z - z^{k+0.5} \rangle \geq 0, \quad \forall z \in \mathcal{Z}.$$

Taking $z = z^{k+1}$, we get

$$(a) \leq \frac{1}{2}\|z^{k+1} - z^k\|^2 - \frac{1}{2}\|z^{k+0.5} - z^k\|^2 - \frac{1}{2}\|z^{k+1} - z^{k+0.5}\|^2. \tag{4.6.40}$$

We can also establish the bound for (b):

$$\begin{aligned}
(b) &= (z^{k+1} - z^{k+0.5})^\top \left(-\alpha \hat{F}(z^{k+0.5}) + \alpha \hat{F}(z^k) - \alpha \hat{F}(z^k) + \eta \hat{F}(z^k) + (\gamma - \beta)(z^k - z^{k-1}) \right) \\
&\leq \alpha \|z^{k+1} - z^{k+0.5}\| \|\hat{F}(z^k) - \hat{F}(z^{k+0.5})\| + (\eta - \alpha)(z^{k+1} - z^{k+0.5})^\top \hat{F}(z^k) \\
&\quad + (\gamma - \beta)(z^{k+1} - z^{k+0.5})^\top (z^k - z^{k-1}).
\end{aligned}$$

Note the following bound from the Lipschitz continuity:

$$\begin{aligned}
\|\hat{F}(z) - \hat{F}(z')\| &= \|\hat{F}(z) - F(z) + F(z') - \hat{F}(z') + F(z) - F(z')\| \\
&\leq \varepsilon_z + \varepsilon_{z'} + L\|z - z'\|,
\end{aligned} \tag{4.6.41}$$

for any $z, z' \in \mathcal{Z}$, where we used the definition of the stochastic error term

$$\varepsilon_z = \left\| F(z) - \hat{F}(z) \right\|. \tag{4.6.42}$$

Therefore,

$$\begin{aligned} & \alpha \|z^{k+1} - z^{k+0.5}\| \|\hat{F}(z^k) - \hat{F}(z^{k+0.5})\| \\ \leq\ & \frac{1}{2}\left(\|z^{k+1} - z^{k+0.5}\|^2 + \alpha^2 \|\hat{F}(z^k) - \hat{F}(z^{k+0.5})\|^2 \right) \\ \leq\ & \frac{1}{2}\|z^{k+1} - z^{k+0.5}\|^2 + \alpha^2(\varepsilon_{z^k} + \varepsilon_{z^{k+0.5}})^2 + \alpha^2 L^2 \|z^k - z^{k+0.5}\|^2. \end{aligned} \tag{4.6.43}$$

Furthermore,

$$\begin{aligned} & (\gamma - \beta)(z^{k+1} - z^{k+0.5})^\top (z^k - z^{k-1}) \\ \leq\ & \frac{1}{2}|\gamma - \beta| \left(\|z^{k+1} - z^{k+0.5}\|^2 + \|z^k - z^{k-1}\|^2 \right) \\ \leq\ & |\gamma - \beta| \left(\|z^{k+1} - z^k\|^2 + \|z^{k+0.5} - z^k\|^2 + \|z^k - z^*\|^2 + \|z^{k-1} - z^*\|^2 \right) \\ \leq\ & |\gamma - \beta| \left(2\|z^{k+1} - z^*\|^2 + 2\|z^k - z^*\|^2 + \|z^{k+0.5} - z^k\|^2 + \|z^k - z^*\|^2 + \|z^{k-1} - z^*\|^2 \right) \\ =\ & |\gamma - \beta| \left(2\|z^{k+1} - z^*\|^2 + 3\|z^k - z^*\|^2 + \|z^{k+0.5} - z^k\|^2 + \|z^{k-1} - z^*\|^2 \right). \end{aligned}$$

The resulting bound for (b) becomes:

$$\begin{aligned} (b) \ \leq\ & \frac{1}{2}\|z^{k+1} - z^{k+0.5}\|^2 + \alpha^2(\epsilon_{z^k} + \epsilon_{z^{k+0.5}})^2 + \left(\alpha^2 L^2 + |\gamma - \beta|\right) \|z^k - z^{k+0.5}\|^2 \\ & + (\eta - \alpha)(z^{k+1} - z^{k+0.5})^\top \hat{F}(z^k) \\ & + |\gamma - \beta| \left(2\|z^{k+1} - z^*\|^2 + 3\|z^k - z^*\|^2 + \|z^{k-1} - z^*\|^2 \right). \end{aligned} \tag{4.6.44}$$

Next let us bound (c) in (4.6.39). We have,

$$\begin{aligned} (c) \ =\ & -\alpha(z^{k+0.5} - z^*)^\top \hat{F}(z^{k+0.5}) + \gamma(z^{k+0.5} - z^*)^\top (z^k - z^{k-1}) \\ \leq\ & -\alpha(z^{k+0.5} - z^*)^\top \hat{F}(z^{k+0.5}) + \frac{1}{2}\gamma \left(\|z^{k+0.5} - z^*\|^2 + \|z^k - z^{k-1}\|^2 \right) \\ \leq\ & -\alpha(z^{k+0.5} - z^*)^\top \hat{F}(z^{k+0.5}) \\ & + \gamma \left(\|z^{k+0.5} - z^k\|^2 + \|z^k - z^*\|^2 + \|z^k - z^*\|^2 + \|z^{k-1} - z^*\|^2 \right) \\ =\ & -\alpha(z^{k+0.5} - z^*)^\top \hat{F}(z^{k+0.5}) + \gamma \left(\|z^{k+0.5} - z^k\|^2 + 2\|z^k - z^*\|^2 + \|z^{k-1} - z^*\|^2 \right). \end{aligned}$$

(4.6.45)

Combining the bounds for $(a), (b), (c)$ from (4.6.40), (4.6.44), and (4.6.45), it follows from

(4.6.39) that

$$
\begin{aligned}
& (z^{k+1}-z^*)^\top\left(-\alpha\hat{F}(z^{k+0.5})+\gamma(z^k-z^{k-1})\right) \\
\leq\ & -\alpha(z^{k+0.5}-z^*)^\top\hat{F}(z^{k+0.5})+(\eta-\alpha)(z^{k+1}-z^{k+0.5})^\top\hat{F}(z^k)+\alpha^2(\varepsilon_{z^k}+\varepsilon_{z^{k+0.5}})^2 \\
& +2\left(|\gamma-\beta|\right)\|z^{k+1}-z^*\|^2+\left(2\gamma+3|\gamma-\beta|\right)\|z^k-z^*\|^2+\left(|\gamma-\beta|+\gamma\right)\|z^{k-1}-z^*\|^2 \\
& +\frac{1}{2}\|z^{k+1}-z^k\|^2+\left(\alpha^2L^2+|\gamma-\beta|+\gamma-\frac{1}{2}\right)\|z^{k+0.5}-z^k\|^2.
\end{aligned}
\tag{4.6.46}
$$

We also need to bound the following term in (4.6.38):

$$
\begin{aligned}
& -\tau(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k-1})\right)\leq\tau\|z^{k+1}-z^*\|\|\hat{F}(z^k)-\hat{F}(z^{k-1})\| \\
\overset{(4.6.41)}{\leq}\ & \tau L\|z^{k+1}-z^*\|\left(\frac{1}{L}(\varepsilon_{z^k}+\varepsilon_{z^{k-1}})+\|z^k-z^{k-1}\|\right) \\
\leq\ & \frac{\tau L}{2}\|z^{k+1}-z^*\|^2+\frac{\tau}{L}(\varepsilon_{z^k}+\varepsilon_{z^{k-1}})^2+\tau L\|z^k-z^{k-1}\|^2 \\
\leq\ & \frac{\tau L}{2}\|z^{k+1}-z^*\|^2+\frac{\tau}{L}(\varepsilon_{z^k}+\varepsilon_{z^{k-1}})^2+2\tau L\|z^k-z^*\|^2+2\tau L\|z^{k-1}-z^*\|^2.
\end{aligned}
\tag{4.6.47}
$$

Combining the bounds in (4.6.46) and (4.6.47) with (4.6.38) and multiplying both sides by 2, we have

$$
\begin{aligned}
& (1-4|\gamma-\beta|-\tau L)\|z^{k+1}-z^*\|^2 \\
\leq\ & (1+4\gamma+6|\gamma-\beta|+4\tau L))\|z^k-z^*\|^2+(2|\gamma-\beta|+2\gamma+4\tau L)\|z^{k-1}-z^*\|^2 \\
& +\left(2\alpha^2L^2+2|\gamma-\beta|+2\gamma-1\right)\|z^{k+0.5}-z^k\|^2+2\alpha^2(\varepsilon_{z^k}+\varepsilon_{z^{k+0.5}})^2+\frac{2\tau}{L}(\varepsilon_{z^k}+\varepsilon_{z^{k-1}})^2 \\
& -2\alpha(z^{k+0.5}-z^*)^\top\hat{F}(z^{k+0.5})+2(\eta-\alpha)(z^{k+1}-z^{k+0.5})^\top\hat{F}(z^k).
\end{aligned}
$$

Let us now take expectation on both sides. Noting $d_{k+1}=\mathbb{E}\left[\|z^{k+1}-z^*\|^2\right]$, $d_k=\mathbb{E}\left[\|z^k-z^*\|^2\right]$, and $d_{k-1}=\mathbb{E}\left[\|z^{k-1}-z^*\|^2\right]$, and noting that $\mathbb{E}[\varepsilon_z^2]\leq\sigma^2$ for all $z\in\mathcal{Z}$ by Assumption (4.1.2), we obtain

$$
\begin{aligned}
& (1-4|\gamma-\beta|-\tau L)d_{k+1} \\
\leq\ & (1+4\gamma+6|\gamma-\beta|+4\tau L)\,d_k+(2|\gamma-\beta|+2\gamma+4\tau L)\,d_{k-1} \\
& +\left(2\alpha^2L^2+2|\gamma-\beta|+2\gamma-1\right)\mathbb{E}\left[\|z^{k+0.5}-z^k\|^2\right]+8\left(\alpha^2+\frac{\tau}{L}\right)\sigma^2 \\
& -2\alpha\mathbb{E}\left[(z^{k+0.5}-z^*)^\top\hat{F}(z^{k+0.5})\right]+2(\eta-\alpha)\mathbb{E}\left[(z^{k+1}-z^{k+0.5})^\top\hat{F}(z^k)\right].
\end{aligned}
\tag{4.6.48}
$$

We shall replace $\hat{F}(z^{k+0.5})$ with $F(z^{k+0.5})+\hat{F}(z^{k+0.5})-F(z^{k+0.5})$ in the last line and note

that

$$
\begin{aligned}
& \mathbb{E}\left[(z^{k+0.5}-z^*)^\top F(z^{k+0.5})\right] \\
= \; & \mathbb{E}\left[(z^{k+0.5}-z^*)^\top \left(F(z^{k+0.5})-F(z^*)\right)+(z^{k+0.5}-z^*)^\top F(z^*)\right] \\
\geq \; & \mathbb{E}\left[\mu\|z^{k+0.5}-z^*\|^2\right] \geq \frac{\mu}{2}d_k - \mu\mathbb{E}\left[\|z^{k+0.5}-z^k\|^2\right],
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbb{E}\left[(z^{k+0.5}-z^*)^\top \left(\hat{F}(z^{k+0.5})-F(z^{k+0.5})\right)\right] \\
= \; & \mathbb{E}\left[\mathbb{E}\left[(z^{k+0.5}-z^*)^\top \left(\hat{F}(z^{k+0.5})-F(z^{k+0.5})\right)|\xi^{[k]}\right]\right] \\
= \; & \mathbb{E}\left[(z^{k+0.5}-z^*)^\top \left(\mathbb{E}\left[\hat{F}(z^{k+0.5})|\xi^{[k]}\right]-F(z^{k+0.5})\right)\right] \\
\geq \; & -\mathbb{E}\left[\|z^{k+0.5}-z^*\|\|\mathbb{E}[\hat{F}(z^{k+0.5})]-F(z^{k+0.5})\|\right] \geq -\mathbb{E}\left[\|z^{k+0.5}-z^*\|\cdot\delta\right] \\
\geq \; & -\mathbb{E}\left[\frac{2\delta^2}{\mu}+\frac{\mu\|z^{k+0.5}-z^*\|^2}{8}\right] \geq -\mathbb{E}\left[\frac{2\delta^2}{\mu}+\frac{\mu\|z^{k+0.5}-z^k\|^2}{4}+\frac{\mu\|z^k-z^*\|^2}{4}\right]
\end{aligned}
$$

Further note that we have denoted $\xi^{[k]} = (\xi^0, \xi^{0.5}, \xi^1, \xi^{1.5}, ..., \xi^{k-0.5}, \xi^k)$ to be the collection of random vectors sampled up until the iterate z^k. Therefore, $z^{k+0.5}$ is a known vector given $\xi^{[k]}$.

Putting the above two bounds back into (4.6.48), we arrive at the desired bound:

$$
\begin{aligned}
& (1-4|\gamma-\beta|-\tau L)d_{k+1} \\
\leq \; & \left(1+4\gamma+6|\gamma-\beta|+4\tau L-\frac{1}{2}\alpha\mu\right)d_k + \left(2|\gamma-\beta|+2\gamma+4\tau L\right)d_{k-1} \\
& +\left(2\alpha^2L^2+2|\gamma-\beta|+2\gamma+\frac{5}{2}\alpha\mu-1\right)\mathbb{E}\left[\|z^{k+0.5}-z^k\|^2\right]+8\left(\alpha^2+\frac{\tau}{L}\right)\sigma^2+\frac{4\alpha\delta^2}{\mu} \\
& +2(\eta-\alpha)\mathbb{E}\left[(z^{k+1}-z^{k+0.5})^\top\hat{F}(z^k)\right].
\end{aligned}
$$

4.6.2 Proof of Theorem 4.2.2

By condition (4.2.8), we have $t_3 < 1$. Let us start with divide both sides of (4.2.7) with $1-t_3$:

$$
\begin{aligned}
d_{k+1} \; & \leq \; \left(1-\frac{t_1-t_3}{1-t_3}\right)d_k + \frac{t_2}{1-t_3}d_{k-1} + \frac{8\left(\alpha^2+\frac{\tau}{L}\right)\sigma^2}{1-t_3} + \frac{4\alpha\delta^2}{\mu(1-t_3)} \\
& = \; (1-a_1)d_k + a_2\cdot d_{k-1} + a_3\cdot\sigma^2 + a_4\cdot\delta^2, \qquad (4.6.49)
\end{aligned}
$$

where we let $(a_1, a_2, a_3, a_4) = \left(\frac{t_1-t_3}{1-t_3}, \frac{t_2}{1-t_3}, \frac{8\left(\alpha^2+\frac{\tau}{L}\right)}{1-t_3}, \frac{4\alpha\delta^2}{\mu(1-t_3)}\right)$ to simplify notations in the rest of this proof. Note that we have $1 > a_1 > a_2$ by condition (4.2.8). It is elementary

to verify that $a_2 \le \left(1 - \frac{a_1 - a_2}{2}\right) \cdot \frac{a_1 + a_2}{2}$, and by rearranging terms in (4.6.49), we have the following

$$\begin{aligned} d_{k+1} + \frac{a_1 + a_2}{2} d_k &\le \left(1 - \frac{a_1 - a_2}{2}\right) d_k + a_2 \cdot d_{k-1} + a_3 \cdot \sigma^2 + a_4 \cdot \delta^2 \\ &\le \left(1 - \frac{a_1 - a_2}{2}\right) \left(d_k + \frac{a_1 + a_2}{2} d_{k-1}\right) + a_3 \cdot \sigma^2 + a_4 \cdot \delta^2 \end{aligned}$$

A recursive argument yields the following result:

$$\begin{aligned} & d_{k+1} + \frac{a_1 + a_2}{2} d_k \\ \le\ & \left(1 - \frac{a_1 - a_2}{2}\right)^{k+1} \left(d_0 + \frac{a_1 + a_2}{2} d_{-1}\right) + \left(a_3 \cdot \sigma^2 + a_4 \cdot \delta^2\right) \cdot \sum_{i=0}^{k} \left(1 - \frac{a_1 - a_2}{2}\right)^i \\ \le\ & \left(1 - \frac{a_1 - a_2}{2}\right)^{k+1} \cdot \frac{2 + a_1 + a_2}{2} \|z^0 - z^*\|^2 + \left(a_3 \cdot \sigma^2 + a_4 \cdot \delta^2\right) \cdot \frac{2}{a_1 - a_2}. \end{aligned}$$

Note that $d_0 = d_{-1} = \|z^0 - z^*\|^2$. The statement in Theorem 4.2.2 follows by letting $q = \frac{2}{a_1 - a_2} = \frac{2(1-t_3)}{t_1 - t_2 - t_3}$.

4.6.3 Proof of Proposition 4.2.3

For the choice of parameters $(\alpha, \beta, \gamma, \eta, \tau) = \left(\frac{1}{4L}, \frac{1}{128\kappa}, \frac{1}{128\kappa}, \frac{1}{4L}, \frac{1}{128L\kappa}\right)$, we have $(t_1, t_2, t_3) = \left(\frac{1}{16\kappa}, \frac{3}{64\kappa}, \frac{1}{128\kappa}\right)$ by the relation (4.2.6). Additionally,

$$2\alpha^2 L^2 + 2|\gamma - \beta| + 2\gamma + \frac{5}{2}\alpha\mu - 1 = \frac{1}{8} + \frac{1}{64\kappa} + \frac{5}{8\kappa} - 1 < 0.$$

Therefore, both conditions (4.2.5) and (4.2.8) are satisfied.

Now, from (4.2.4), we have

$$\begin{aligned} \left(1 - \frac{1}{128\kappa}\right) d_{k+1} &\le \left(1 - \frac{1}{16\kappa}\right) d_k + \frac{3}{64\kappa} d_{k-1} + 8\left(\frac{1}{16L^2} + \frac{1}{128L^2\kappa}\right)\sigma^2 + \frac{\delta^2}{L\mu} \\ &\le \left(1 - \frac{1}{16\kappa}\right) d_k + \frac{3}{64\kappa} d_{k-1} + \frac{9\sigma^2}{16L^2} + \frac{\delta^2}{L\mu}. \end{aligned} \tag{4.6.50}$$

Divide both sides with $1 - \frac{1}{128\kappa}$ and note that $\left(1 - \frac{1}{128\kappa}\right)^{-1} \le \frac{128}{127}$, we have:

$$\begin{aligned} d_{k+1} &\le \frac{1 - \frac{1}{16\kappa}}{1 - \frac{1}{128\kappa}} d_k + \frac{6}{127\kappa} d_{k-1} + \frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu} \\ &= \left(1 - \frac{\frac{7}{128\kappa}}{1 - \frac{1}{128\kappa}}\right) d_k + \frac{6}{127\kappa} d_{k-1} + \frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu} \\ &\le \left(1 - \frac{7}{128\kappa}\right) d_k + \frac{6}{127\kappa} d_{k-1} + \frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu}. \end{aligned} \tag{4.6.51}$$

We can move a part of d_k to the LHS and form the following:

$$
\begin{aligned}
(4.6.52) \qquad & d_{k+1} + \frac{13}{256\kappa} d_k \le \left(1 - \frac{1}{256\kappa}\right) d_k + \frac{6}{127\kappa} d_{k-1} + \frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu} \\
\le \quad & \left(1 - \frac{1}{256\kappa}\right)\left(d_k + \frac{13}{256\kappa} d_{k-1}\right) + \frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu} \\
\le \quad & \left(1 - \frac{1}{256\kappa}\right)^{k+1}\left(d_0 + \frac{13}{256\kappa} d_{-1}\right) + \left(\frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu}\right) \cdot \sum_{i=0}^{k} \left(1 - \frac{1}{256\kappa}\right)^{i} \\
\le \quad & \left(1 - \frac{1}{256\kappa}\right)^{k+1} \cdot \frac{269}{256}\|z^0 - z^*\|^2 + \left(\frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu}\right) \cdot \frac{1 - \left(1 - \frac{1}{256\kappa}\right)^{k+1}}{\frac{1}{256\kappa}} \\
\le \quad & \left(1 - \frac{1}{256\kappa}\right)^{k+1} \cdot \frac{269}{256}\|z^0 - z^*\|^2 + \left(\frac{72\sigma^2}{127L^2} + \frac{128\delta^2}{127L\mu}\right) \cdot 256\kappa.
\end{aligned}
$$

Note that the second inequality is due to $\left(1 - \frac{1}{256\kappa}\right)^{-1} \frac{6}{127\kappa} \le \frac{256}{255} \cdot \frac{6}{127\kappa} \le \frac{13}{256\kappa}$, and we also have $d_0 = d_{-1} = \|z^0 - z^*\|^2$ for the fourth inequality. Finally, the LHS of the above inequality can be lower bounded by d_{k+1}, thus completing the proof.

4.6.4 Proof of Lemma 4.2.4

We start by using the 1-co-coerciveness of the projection operator $P_{\mathcal{Z}}(\cdot)$:

$$
\begin{aligned}
& \|z^{k+1} - z^*\|^2 \\
= \quad & \|P_{\mathcal{Z}}\left(z^k - \alpha\hat{F}(z^k) + \gamma(z^k - z^{k-1}) - \tau\left(\hat{F}(z^k) - \hat{F}(z^{k-1})\right)\right) - P_{\mathcal{Z}}\left(z^* - \alpha F(z^*)\right)\|^2 \\
\le \quad & (z^{k+1} - z^*)^\top \left(z^k - \alpha\hat{F}(z^k) + \gamma(z^k - z^{k-1}) - \tau\left(\hat{F}(z^k) - \hat{F}(z^{k-1})\right) - (z^* - \alpha F(z^*))\right) \\
= \quad & (z^{k+1} - z^*)^\top \left((z^k - z^*) - \alpha\left(\hat{F}(z^k) - F(z^*)\right) + \gamma(z^k - z^{k-1}) - \tau\left(\hat{F}(z^k) - \hat{F}(z^{k-1})\right)\right).
\end{aligned}
\tag{4.6.53}
$$

Next, let us bound the above four terms separately:

$$
(z^{k+1} - z^*)^\top (z^k - z^*) = \frac{1}{2}\left(\|z^{k+1} - z^*\|^2 + \|z^k - z^*\|^2 - \|z^{k+1} - z^k\|^2\right),
$$

and

$$
(z^{k+1} - z^*)^\top \left(\hat{F}(z^k) - F(z^*)\right) \quad = \quad (z^{k+1} - z^*)^\top \left(\hat{F}(z^k) - \hat{F}(z^{k+1}) + \hat{F}(z^{k+1}) - F(z^*)\right)
$$

where

$$
\begin{aligned}
& (z^{k+1} - z^*)^\top \left(\hat{F}(z^{k+1}) - F(z^*)\right) \\
= \quad & (z^{k+1} - z^*)^\top \left(\hat{F}(z^{k+1}) - F(z^{k+1}) + F(z^{k+1}) - F(z^*)\right) \\
\ge \quad & (z^{k+1} - z^*)^\top \left(\hat{F}(z^{k+1}) - F(z^{k+1})\right) + \mu\|z^{k+1} - z^*\|^2,
\end{aligned}
$$

and

$$(z^{k+1}-z^*)^\top(z^k-z^{k-1}) \quad \leq \quad \frac{\gamma}{2}\left(\|z^{k+1}-z^*\|^2+\|z^k-z^{k-1}\|^2\right),$$

and

$$\begin{aligned}
& -\tau(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k-1})\right) = -\tau(z^{k+1}-z^k+z^k-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k-1})\right) \\
\leq \quad & \tau\|z^{k+1}-z^k\|\|\hat{F}(z^k)-\hat{F}(z^{k-1})\| - \tau(z^k-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k-1})\right) \\
\leq \quad & \frac{1}{4}\|z^{k+1}-z^k\|^2+\tau^2\|\hat{F}(z^k)-\hat{F}(z^{k-1})\|^2 - \tau(z^k-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k-1})\right),
\end{aligned}$$

where

$$\tau^2\|\hat{F}(z^k)-\hat{F}(z^{k-1})\|^2 \quad \overset{(4.6.41)}{\leq} \quad 2\tau^2(\varepsilon_{z^k}+\varepsilon_{z^{k-1}})^2+2\tau^2L^2\|z^k-z^{k-1}\|^2. \tag{4.6.54}$$

Putting the above bounds back to (4.6.53), we get:

$$\begin{aligned}
& \left(\frac{1}{2}+\alpha\mu-\frac{\gamma}{2}\right)\|z^{k+1}-z^*\|^2+\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k+1})\right)+\frac{1}{4}\|z^{k+1}-z^k\|^2 \\
\leq \quad & \frac{1}{2}\|z^k-z^*\|^2+\tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^k-z^{k-1}\|^2 \\
& +2\tau^2(\varepsilon_{z^k}+\varepsilon_{z^{k-1}})^2-\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^{k+1})-F(z^{k+1})\right).
\end{aligned}$$

Taking expectation on both sides gives us

$$\begin{aligned}
& \mathbb{E}\left[\left(\frac{1}{2}+\alpha\mu-\frac{\gamma}{2}\right)\|z^{k+1}-z^*\|^2+\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k+1})\right)+\frac{1}{4}\|z^{k+1}-z^k\|^2\right] \\
\leq \quad & \mathbb{E}\left[\frac{1}{2}\|z^k-z^*\|^2+\tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^k-z^{k-1}\|^2\right] \\
& +8\tau^2\sigma^2-\mathbb{E}\left[\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^{k+1})-F(z^{k+1})\right)\right].
\end{aligned} \tag{4.6.55}$$

Note that

$$\begin{aligned}
& -\mathbb{E}\left[\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^{k+1})-F(z^{k+1})\right)\right] \\
= \quad & -\alpha\mathbb{E}\left[\mathbb{E}\left[(z^{k+1}-z^*)^\top\left(\hat{F}(z^{k+1})-F(z^{k+1})\right)|\xi^{[k]}\right]\right] \\
= \quad & -\alpha\mathbb{E}\left[(z^{k+1}-z^*)^\top\left(\mathbb{E}\left[\hat{F}(z^{k+1})|\xi^{[k]}\right]-F(z^{k+1})\right)\right] \\
\leq \quad & \alpha\mathbb{E}\left[\|z^{k+1}-z^*\|\|\mathbb{E}[\hat{F}(z^{k+1})]-F(z^{k+1})\|\right] \\
\leq \quad & \alpha\mathbb{E}\left[\delta\|z^{k+1}-z^*\|\right] \leq \alpha\mathbb{E}\left[\frac{\delta^2}{2\mu}+\frac{\mu}{2}\|z^{k+1}-z^*\|^2\right].
\end{aligned} \tag{4.6.56}$$

Here we define $\xi^{[k]} = (\xi^0, \xi^1, ..., \xi^k)$ to be the collection of random vectors sampled up until the iterate z^k, and z^{k+1} is known given $\xi^{[k]}$.

Therefore, (4.6.55) becomes

$$\begin{aligned}
& \mathbb{E}\left[\left(\frac{1}{2}+\frac{\alpha\mu}{2}-\frac{\gamma}{2}\right)\|z^{k+1}-z^*\|^2+\alpha(z^{k+1}-z^*)^\top\left(\hat{F}(z^k)-\hat{F}(z^{k+1})\right)+\frac{1}{4}\|z^{k+1}-z^k\|^2\right] \\
\leq\ & \mathbb{E}\left[\frac{1}{2}\|z^k-z^*\|^2+\tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)+\left(2\tau^2L^2+\frac{\gamma}{2}\right)\|z^k-z^{k-1}\|^2\right] \\
& +8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu},
\end{aligned}$$

completing the proof.

4.6.5 Proof of Theorem 4.2.5

Continuing from (4.2.14), we have:

$$\begin{aligned}
V_k &\leq \left(1+\frac{\theta}{\kappa}\right)^{-k}V_0+\sum_{i=1}^{k}\left(1+\frac{\theta}{\kappa}\right)^{-i}\cdot\left(8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu}\right) \\
&= \frac{1}{2}\left(1+\frac{\theta}{\kappa}\right)^{-k}\|z^0-z^*\|^2+\frac{1-\left(1+\frac{\theta}{\kappa}\right)^{-k}}{\frac{\theta}{\kappa}}\cdot\left(8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu}\right),
\end{aligned}$$

where we use $z^{-1}=z^0$ for V_0.

Finally, with the following bound:

$$\begin{aligned}
& \tau(z^k-z^*)^\top\left(\hat{F}(z^{k-1})-\hat{F}(z^k)\right)\geq-\tau\|z^k-z^*\|\|\hat{F}(z^{k-1})-\hat{F}(z^k)\| \\
\geq\ & -\frac{1}{4}\|z^k-z^*\|^2-\tau^2\|\hat{F}(z^{k-1})-\hat{F}(z^k)\|^2 \\
\overset{(4.6.41)}{\geq}\ & -\frac{1}{4}\|z^k-z^*\|^2-\tau^2\left(2L^2\|z^{k-1}-z^k\|^2+2(\varepsilon_{z^{k-1}}+\varepsilon_{z^k})^2\right),
\end{aligned}$$

we can lower bound V_k as $\frac{1}{4}\mathbb{E}\left[\|z^k-z^*\|^2\right]-8\tau^2\sigma^2\leq V_k$. Therefore

$$\begin{aligned}
& \mathbb{E}\left[\|z^k-z^*\|^2\right] \\
\leq\ & 2\left(1+\frac{\theta}{\kappa}\right)^{-k}\|z^0-z^*\|^2+\frac{4\kappa}{\theta}\cdot\left(1-\left(1+\frac{\theta}{\kappa}\right)^{-k}\right)\cdot\left(8\tau^2\sigma^2+\frac{\alpha\delta^2}{2\mu}\right)+32\tau^2\sigma^2 \\
\leq\ & 2\left(1+\frac{\theta}{\kappa}\right)^{-k}\|z^0-z^*\|^2+\left(\frac{\kappa}{\theta}+1\right)\cdot 32\tau^2\sigma^2+\frac{2\kappa\alpha\delta^2}{\theta\mu}.
\end{aligned}$$

(4.6.57)

The statement in Theorem 4.2.5 follows by noting $\left(1+\frac{\theta}{\kappa}\right)^{-1}=1-\frac{\theta}{\kappa+\theta}$.

4.6.6 Proof of Lemma 4.3.2

We will derive the first bound in (4.3.26); the second bound is similar and will be omitted. Notice that

$$\begin{aligned}
& \mathbb{E}_{\xi,u}\left[\|F_{\rho_x}(x,y,\xi,u)\|^2\right] = \mathbb{E}_{\xi}\left[\mathbb{E}_u\left[\|F_{\rho_x}(x,y,\xi,u)\|^2\right]\right] \\
\overset{(4.3.23)}{\leq} \quad & \mathbb{E}_{\xi}\left[2n\|\nabla_x \hat{f}(x,y,\xi)\|^2\right] + \frac{\rho_x^2 L^2 n^2}{2} \\
= \quad & 2n\left[\mathbb{E}_{\xi}\left[\|\nabla_x f(x,y)\|^2 + 2\nabla_x f(x,y)^\top\left(\nabla_x \hat{f}(x,y,\xi) - \nabla_x f(x,y)\right)\right.\right. \\
& \left.\left. +\|\nabla_x \hat{f}(x,y,\xi) - \nabla_x f(x,y)\|^2\right]\right] + \frac{\rho_x^2 L^2 n^2}{2} \\
\overset{(4.3.21)}{=} \quad & 2n\left(\|\nabla_x f(x,y)\|^2 + \mathbb{E}_{\xi}\left[\|\nabla_x \hat{f}(x,y,\xi) - \nabla_x f(x,y)\|^2\right]\right) + \frac{\rho_x^2 L^2 n^2}{2} \\
\leq \quad & 2n\left(M^2 + \sigma^2\right) + \frac{\rho_x^2 L^2 n^2}{2}.
\end{aligned}$$

Further note that

$$\begin{aligned}
& \mathbb{E}_{\xi,u}\left[\|F_{\rho_x}(x,y,\xi,u) - \nabla_x f_{\rho_x}(x,y)\|^2\right] \\
= \quad & \mathbb{E}_{\xi,u}\left[\|F_{\rho_x}(x,y,\xi,u)\|^2 - 2F_{\rho_x}(x,y,\xi,u)^\top \nabla_x f_{\rho_x}(x,y) + \|\nabla_x f_{\rho_x}(x,y)\|^2\right] \\
\overset{(4.3.25)}{=} \quad & \mathbb{E}_{\xi,u}\left[\|F_{\rho_x}(x,y,\xi,u)\|^2 - \|\nabla_x f_{\rho_x}(x,y)\|^2\right] \leq \mathbb{E}_{\xi,u}\left[\|F_{\rho_x}(x,y,\xi,u)\|^2\right] \\
\leq \quad & 2n(M^2 + \sigma^2) + \frac{\rho_x^2 L^2 n^2}{2},
\end{aligned}$$

completing the proof for (4.3.26).

4.6.7 Proof of Lemma 4.3.3

The logic line of the proof for this lemma is very similar to the proof in Appendix 4.6.1, with the stochastic mapping $\hat{F}(z^k)$ replaced by the stochastic zeroth-order gradient $\hat{F}_\rho^k(z^k)$. Therefore, we shall refrain from repeating similar analysis, but highlight the main differences instead, which lies in bounding $\|F_\rho(z) - F(z)\|$ and $\|\hat{F}_\rho^k(z^k) - \hat{F}_\rho^{k'}(z^{k'})\|$. The former can be bounded by $\frac{L\sqrt{\rho_x^2 n^2 + \rho_y^2 m^2}}{2}$, which is a direct result from (4.3.22). We illustrate the bound on the latter in the following derivations. First, for $F(z) = \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}$, we shall have

$$\|F(z) - F(z')\| \leq L\|z - z'\|, \quad \forall z, z' \in \mathcal{Z} \tag{4.6.58}$$

where $L = 2 \cdot \max(L_x, L_y, L_{xy})$, because

$$\begin{aligned}
& \|F(z) - F(z')\|^2 = \|\nabla_x f(x,y) - \nabla_x f(x',y')\|^2 + \|\nabla_y f(x,y) - \nabla_y f(x',y')\|^2 \\
\leq \quad & 2L_x^2\|x - x'\|^2 + 2L_{xy}^2\|y - y'\|^2 + 2L_y^2\|y - y'\|^2 + 2L_{xy}^2\|x - x'\|^2 \\
\leq \quad & L^2(\|x - x'\|^2 + \|y - y'\|^2) = L^2\|z - z'\|^2.
\end{aligned}$$

Next, by denoting $\varepsilon_{z^k} = \|\hat{F}^k_\rho(z^k) - F_\rho(z^k)\|$, we have

$$
\begin{aligned}
(4.6.59) \qquad & \|\hat{F}^k_\rho(z^k) - \hat{F}^{k'}_\rho(z^{k'})\| \\
= \quad & \|\hat{F}^k_\rho(z^k) - F_\rho(z^k) + F_\rho(z^{k'}) - \hat{F}^{k'}_\rho(z') + F_\rho(z^k) - F_\rho(z^{k'})\| \\
\leq \quad & \varepsilon_{z^k} + \varepsilon_{z^{k'}} + \|F_\rho(z^k) - F_\rho(z^{k'})\| \\
= \quad & \varepsilon_{z^k} + \varepsilon_{z^{k'}} + \|F_\rho(z^k) - F(z^k) + F(z^{k'}) - F_\rho(z^{k'}) + F(z^k) - F(z^{k'})\| \\
\overset{(4.3.22),(4.6.58)}{\leq} \quad & \varepsilon_{z^k} + \varepsilon_{z^{k'}} + L\sqrt{\rho_x^2 n^2 + \rho_y^2 m^2} + L\|z^k - z^{k'}\|.
\end{aligned}
$$

Now, similar bounds as in (4.6.43) and (4.6.47) can be constructed by taking $k' = k + 0.5$ and $k' = k - 1$ respectively in (4.6.59). We finally note that $\mathbb{E}[\epsilon^2_{z^k}] \leq \frac{2\tilde{\sigma}^2}{t_k}$ to proceed through the rest of the proof. The interested readers are referred to Appendix A.7 of the archival report version [26] of the current paper.

4.6.8 Proof of Proposition 4.3.4

Let us present the following lemma first, then the results of Proposition (4.3.4) shall follow immediately. A similar proof can be found in [44], Lemma 11.

Lemma 4.6.1. *Let $K > 1$ be the total iteration number, $\{V_k\}_{k=0,..,K}$ and $\{t_k\}_{k=0,..,K}$ be positive sequences, $\frac{1}{4} < C < 1$ and $a > 0$ be constants, such that the following holds:*

$$
V_k \leq C \cdot V_{k-1} + \frac{aC}{t_{k-1}}, \tag{4.6.60}
$$

for $k = 1, ..., K$. If we take $t_k = C^{-k}$, then we further have:

$$
V_K \leq \left(\frac{1}{2}\right)^{\frac{K}{b}} \left(\frac{1}{2} \cdot V_0 + \frac{4a}{1-C}\right), \tag{4.6.61}
$$

where $b = \lceil \log_C \frac{1}{4} \rceil > 1$.

Proof. First note that $C^b \leq \frac{1}{4}$. Using $t_k = C^{-k}$ we have:

$$
V_K \quad \leq \quad C \cdot V_{K-1} + a\frac{C}{t_{K-1}} \leq C^b \cdot V_{K-b} + a\sum_{k=1}^{b} \frac{C^k}{t_{K-k}} = C^b \cdot V_{K-b} + ab \cdot C^K,
$$

which holds for $K \geq b$. Without loss of generality, assume $(K \mod b) = 0$, then the following holds:

$$
\begin{aligned}
b \cdot C^K \quad &= \quad C \cdot C^{K-1} + C^2 \cdot C^{K-2} + \cdots + C^b \cdot C^{K-b} \\
&\leq \quad C \cdot \frac{C^{K-2b}}{4} + C^2 \cdot \frac{C^{K-2b}}{4} + \cdots + C^b \cdot \frac{C^{K-2b}}{4} = \frac{1-C^b}{1-C} \cdot \frac{C^{K-2b}}{4} \leq \frac{C^{K-2b}}{4(1-C)}.
\end{aligned}
$$

The first inequality is due to $C^{K-b+i} \leq C^{K-b} = C^{K-2b} \cdot C^b \leq \frac{C^{K-2b}}{4}$ for any $i = 1, 2, \ldots$. Therefore we have:

$$\begin{aligned}
V_K &\leq C^b \cdot V_{K-1} + a(b \cdot C^K) \leq \frac{1}{4}\left(V_{K-b} + \frac{a}{1-C} \cdot C^{K-2b}\right) \\
&\leq \frac{1}{2}\left(V_{K-b} + \frac{a}{1-C} \cdot C^{K-2b}\right) \\
&\leq \frac{1}{2}\left(C^b \cdot V_{K-2b} + a(b \cdot C^{K-b}) + \frac{a}{1-C} \cdot C^{K-2b}\right) \\
&\leq \frac{1}{2}\left(\frac{1}{4} \cdot V_{K-2b} + \frac{1}{4} \cdot \frac{a}{1-C} \cdot C^{K-3b} + \frac{1}{4} \cdot \frac{a}{1-C} \cdot C^{K-3b}\right) \\
&\leq \left(\frac{1}{2}\right)^2 \left(V_{K-2b} + \frac{a}{1-C} \cdot C^{K-3b}\right) \leq \left(\frac{1}{2}\right)^{\frac{K}{b}-1} \left(V_b + \frac{a}{1-C}\right) \\
&\leq \left(\frac{1}{2}\right)^{\frac{K}{b}-1} \left(C^b \cdot V_0 + a(b \cdot C^b) + \frac{a}{1-C}\right).
\end{aligned}$$

Since $C < 1$, obviously we have $b \cdot C^b \leq \sum_{i=1}^{b} C^i \leq \frac{1}{1-C}$. Therefore,

$$\begin{aligned}
V_K &\leq \left(\frac{1}{2}\right)^{\frac{K}{b}-1} \left(C^b \cdot V_0 + a(b \cdot C^b) + \frac{a}{1-C}\right) \leq \left(\frac{1}{2}\right)^{\frac{K}{b}-1} \left(\frac{1}{4} \cdot V_0 + \frac{2a}{1-C}\right) \\
&\leq \left(\frac{1}{2}\right)^{\frac{K}{b}} \left(\frac{1}{2} \cdot V_0 + \frac{4a}{1-C}\right).
\end{aligned}$$

□ □

Now, to prove Proposition 4.3.4, we need to show how the parameter choice in this proposition together with (4.3.29) result in (4.6.60). Let us first bound the variance term by observing $t_k = t_{k+0.5}$ and $\frac{1}{t_{k-1}} = \frac{1}{t_k}\left(1 - \frac{1}{256\kappa}\right)^{-1} \leq \frac{2}{t_k}$:

$$16\tilde{\sigma}^2 \left(\left(\alpha^2 + \frac{\tau}{L}\right)\frac{1}{t_k} + \frac{\alpha^2}{t_{k+0.5}} + \frac{\tau}{L t_{k-1}}\right) \leq \left(\alpha^2 + \frac{\tau}{L}\right)\frac{48\tilde{\sigma}^2}{t_k} \leq \frac{27\tilde{\sigma}^2}{8L^2 t_k}.$$

Denote $C = 1 - \frac{1}{256\kappa}$. The bias term in (4.3.29) can be written us:

$$4\left(\tau L + \alpha^2 L^2 + \frac{\alpha L^2}{4\mu}\right)(\rho_x^2 n^2 + \rho_y^2 m^2) = \left(\tau L + \alpha^2 L^2 + \frac{\alpha L^2}{4\mu}\right)\frac{4C^{2K}}{\kappa^2} \leq \frac{17C^K}{32\kappa}.$$

Substituting the rest of the parameters,we get:

$$\left(1 - \frac{1}{128\kappa}\right) d_{k+1} \leq \left(1 - \frac{1}{16\kappa}\right) d_k + \frac{3}{64\kappa} d_{k-1} + \frac{27\tilde{\sigma}^2}{8L^2 t_k} + \frac{17C^K}{32\kappa}.$$

Since the above inequality is only different from (4.6.50) by the last two terms, following the same procedure as in (4.6.51) and (4.6.52) we immediately obtain:

$$d_{k+1} + \frac{13}{256\kappa} d_k \quad \leq \quad \left(1 - \frac{1}{256\kappa}\right) \cdot \left(d_k + \frac{13}{256\kappa} d_{k-1}\right) + \frac{432\tilde{\sigma}^2}{127L^2 t_k} + \frac{68C^K}{127\kappa},$$

which is in the form of (4.6.60) for $V_k = d_k + \frac{13}{256\kappa} d_{k-1}$, $C = 1 - \frac{1}{256\kappa}$, $a = \frac{432\tilde{\sigma}^2}{127L^2}$, with an additional last term. Lemma 4.6.1 indicates:

$$d_k \leq V_K \leq \left(\frac{1}{2}\right)^{\frac{K}{b}} \left(\frac{1}{2} \cdot V_0 + \frac{4a}{1-C}\right) + \frac{68C^K}{127\kappa} = \left(\frac{1}{2}\right)^{\frac{K}{b}} \mathcal{O}\left(\|z^0 - z^*\|^2 + \frac{\tilde{\sigma}^2}{L\mu}\right) + \mathcal{O}\left(\frac{C^K}{\kappa}\right).$$

By the relation

$$b \cdot \log_2 \frac{1}{\epsilon} \geq \left(\log_C \frac{1}{4}\right) \log_2 \frac{1}{\epsilon} = (\log_{C^{-1}} 4) \log_2 \frac{1}{\epsilon} = 2 \log_{C^{-1}} \frac{1}{\epsilon},$$

the total iteration K to guarantee $d_K \leq \epsilon$ is given by:

$$K \geq \mathcal{O}\left(\max\left\{\log_{C^{-1}} \frac{\|z^0 - z^*\|^2 + \frac{\tilde{\sigma}^2}{L\mu}}{\epsilon}, \log_{C^{-1}} \frac{1}{\kappa\epsilon}\right\}\right) = \mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$$

with the sample complexity as $2 \sum_{k=0}^{K-1} t_k = 2 \sum_{k=0}^{K-1} C^{-k} = \mathcal{O}\left(\frac{\kappa\|z^0 - z^*\|^2}{\epsilon} + \frac{\tilde{\sigma}^2}{\epsilon\mu^2}\right)$.

4.6.9 Proof of Lemma 4.3.5

With the similar logic to the proof in Appendix 4.6.4, we shall focus on the main differences between the two proofs. The first one lies in constructing a similar proof as in (4.6.54), with the help from (4.6.59) by taking $k' = k - 1$. The next one is to construct a similar bound as in (4.6.56), where we shall instead use the bound $\|F_\rho(z) - F(z)\| \leq \frac{L\sqrt{\rho_x^2 n^2 + \rho_y^2 m^2}}{2}$ as a result of (4.3.22). A more detailed proof can be found in Appendix A.9 of the archival report version [26] of the current paper.

4.6.10 Proof of Proposition 4.3.6

Let us first identify (4.3.33) in the form of (4.6.60) and the results shall follow by Lemma 4.6.1. Let $C = \left(1 + \frac{1}{8\kappa}\right)^{-1} = \left(1 - \frac{1}{8\kappa+1}\right)$ and note that $\frac{1}{t_{k-1}} = \frac{1}{Ct_k} \leq \frac{2}{t_k}$, we have

$$16\tau^2\tilde{\sigma}^2 \left(\frac{1}{t_k} + \frac{1}{t_{k-1}}\right) \leq \frac{48\tau^2\sigma^2}{t_k}$$

and

$$L^2 \left(4\tau^2 + \frac{\alpha}{8\mu}\right) (\rho_x^2 n^2 + \rho_y^2 m^2) = \mu^2 \left(4\tau^2 + \frac{\alpha}{8\mu}\right) \cdot C^K.$$

Then we have from (4.3.33):

$$V_k \leq C \cdot V_{k-1} + \frac{48\tau^2\tilde{\sigma}^2 C}{t_{k-1}} + \mu^2\left(4\tau^2 + \frac{\alpha}{8\mu}\right) \cdot C^K.$$

With $a = 48\tau^2\tilde{\sigma}^2$, Lemma 4.6.1 leads to:

$$V_K \leq \left(\frac{1}{2}\right)^{\frac{K}{b}} \left(\frac{1}{2} \cdot V_0 + \frac{4a}{1-C}\right) + \mu^2\left(4\tau^2 + \frac{\alpha}{8\mu}\right) \cdot C^K.$$

Now, let us lower bound V_k by observing:

$$\begin{aligned} & \tau(z^k - z^*)^\top \left(\hat{F}_\rho^{k-1}(z^{k-1}) - \hat{F}_\rho^k(z^k)\right) \geq -\tau\|z^k - z^*\|\|\hat{F}_\rho^{k-1}(z^{k-1}) - \hat{F}_\rho^k(z^k)\| \\ \geq \quad & -\frac{1}{4}\|z^k - z^*\|^2 - \tau^2\|\hat{F}_\rho^{k-1}(z^{k-1}) - \hat{F}_\rho^k(z^k)\|^2 \\ \overset{(4.6.59)}{\geq} \quad & -\frac{1}{4}\|z^k - z^*\|^2 - \tau^2\left(2L^2\|z^{k-1} - z^k\|^2 + 4(\varepsilon_{z^{k-1}} + \varepsilon_{z^k})^2 + 4L^2(\rho_x^2 n^2 + \rho_y^2 m^2)\right. . \end{aligned}$$

Then we have

$$\begin{aligned} V_K \quad \geq \quad & \frac{1}{4}d_K - 16\tau^2\tilde{\sigma}^2\left(\frac{1}{t_K} + \frac{1}{t_{K-1}}\right) - \frac{4\tau^2L^2C^K}{\kappa^2} \geq \frac{1}{4}d_K - 48\tau^2\tilde{\sigma}^2C^K - \frac{4\tau^2L^2C^K}{\kappa^2} \\ = \quad & \frac{1}{4}d_K - aC^K - 4\mu^2\tau^2C^K. \end{aligned}$$

Combining the above results and substituting the parameters gives us:

$$\begin{aligned} \frac{1}{4}d_K \quad \leq \quad & \left(\frac{1}{2}\right)^{\frac{K}{b}} \left(\frac{1}{2} \cdot V_0 + \frac{4a}{1-C}\right) + \left(8\mu^2\tau^2 + \frac{\mu\alpha}{8} + a\right) \cdot C^K \\ = \quad & \left(\frac{1}{2}\right)^{\frac{K}{b}} \mathcal{O}\left(\|z^0 - z^*\|^2 + \frac{\tilde{\sigma}^2}{L\mu}\right) + C^K\mathcal{O}\left(\frac{1}{\kappa^2} + \frac{1}{\kappa} + \frac{\tilde{\sigma}^2}{L\mu}\right). \end{aligned}$$

Following the same logic as in the last part of Appendix 4.6.8, we conclude that with $K = \mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$, we have $d_K \leq \epsilon$ and the sample complexity is

$$\sum_{k=0}^{K-1} t_k = \sum_{k=0}^{K-1} C^{-k} = \mathcal{O}\left(\frac{\kappa\|z^0 - z^*\|^2}{\epsilon} + \frac{\tilde{\sigma}^2}{\epsilon\mu^2}\right).$$

4.7 Appendix B: Proof of the uniform sublinear convergence of the stochastic extra-point method

In order to establish a uniform sublinear convergence, we have to consider parameters that are diminishing with iteration number k. Let us return to the one-iteration relation (4.2.4) and consider the following choice of parameters:

$$(\alpha^{(k)}, \beta^{(k)}, \gamma^{(k)}, \eta^{(k)}, \tau^{(k)}) = \left(\frac{2}{(k+2)\mu}, \frac{\alpha^2\mu^2}{128}, \frac{\alpha^2\mu^2}{128}, \frac{2}{(k+2)\mu}, \frac{\alpha^2\mu}{128\kappa}\right),$$

where we omit the superscript (k) of α on the RHS for notation simplicity. We shall note that here $\alpha = \alpha^{(k)}$ which is dependent on iteration k and follow the same simplification for other parameters throughout the rest of the proof in this appendix unless noted otherwise.

By using the fact $2\alpha \leq \frac{1}{\mu} + \alpha^2\mu$, we have:

$$\begin{aligned}
& \left(2\alpha^2 L^2 + 2|\gamma - \beta| + 2\gamma + 2\alpha\mu - 1\right) \mathbb{E}\left[\|z^{k+0.5} - z^k\|^2\right] \\
\leq\ & \left(2\alpha^2 L^2 + 2\gamma + \alpha^2\mu\right) \mathbb{E}\left[\|z^{k+0.5} - z^k\|^2\right] \\
\leq\ & \alpha^2 \left(2L^2 + \frac{\mu}{64} + \mu\right) \mathbb{E}\left[\|z^{k+0.5} - z^k\|^2\right] \\
\leq\ & \alpha^2 \left(2L^2 + \frac{\mu^2}{64} + \mu^2\right) D^2,
\end{aligned}$$

where in the last inequality we use the boundedness of the feasible set.

Therefore, we could rewrite (4.2.4) into:

$$\begin{aligned}
& (1 - \tau L) d_{k+1} \\
\leq\ & (1 + 4\gamma + 4\tau L - \alpha\mu)\, d_k + (2\gamma + 4\tau L)\, d_{k-1} \\
& + \alpha^2 \left(2L^2 + \frac{\mu^2}{64} + \mu^2\right) D^2 + 8\left(\alpha^2 + \frac{\tau}{L}\right)\sigma^2 + 2\alpha\delta D \\
=\ & (1 + 4\gamma + 4\tau L - \alpha\mu)\, d_k + (2\gamma + 4\tau L)\, d_{k-1} \\
& + \frac{4}{(k+2)^2} \cdot \underbrace{\left(2\kappa^2 D^2 + \frac{D^2}{64} + D^2 + \frac{8\sigma^2}{\mu^2} + \frac{\sigma^2}{128L^2}\right)}_{G} + \frac{4\delta D}{(k+2)\mu}.
\end{aligned}$$

Substituting the parameters with their respective values in the rest of the terms:

$$\begin{aligned}
& \left(1 - \frac{1}{32(k+2)}\right) d_{k+1} \\
\leq\ & \left(1 - \frac{1}{32(k+2)^2}\right) d_{k+1} \\
\leq\ & \left(1 + \frac{1}{8(k+2)^2} + \frac{1}{8(k+2)^2} - \frac{2}{k+2}\right) d_k \\
& + \left(\frac{1}{16(k+2)^2} + \frac{1}{8(k+2)^2}\right) d_{k-1} + \frac{4G}{(k+2)^2} + \frac{4\delta D}{(k+2)\mu} \\
\leq\ & \left(1 - \frac{7}{4(k+2)}\right) d_k + \frac{3}{16(k+2)} d_{k-1} + \frac{4G}{(k+2)^2} + \frac{4\delta D}{(k+2)\mu}.
\end{aligned}$$

Dividing both sides by $1-\frac{1}{32(k+2)}$, and noting that $\left(1-\frac{1}{32(k+2)}\right)^{-1} \leq \frac{32}{31}$, it follows that

$$\begin{aligned}
d_{k+1} &\leq \frac{1-\frac{7}{4(k+2)}}{1-\frac{1}{32(k+2)}} d_k + \frac{6}{31(k+2)} d_{k-1} + \frac{128G}{31(k+2)^2} + \frac{128\delta D}{31(k+2)\mu} \\
&= \left(1 - \frac{\frac{55}{32(k+2)}}{1-\frac{1}{32(k+2)}}\right) d_k + \frac{6}{31(k+2)} d_{k-1} + \frac{128G}{31(k+2)^2} + \frac{128\delta D}{31(k+2)\mu} \\
&\leq \left(1 - \frac{55}{32(k+2)}\right) d_k + \frac{7}{32(k+2)} d_{k-1} + \frac{128G}{31(k+2)^2} + \frac{128\delta D}{31(k+2)\mu}.
\end{aligned}$$

From the above one-iteration inequality, we shall claim the following inequality

$$d_k \leq \frac{Q}{k+2} + \frac{256\delta D}{93\mu}, \quad \forall k \geq 0,$$

where

$$Q = \max\left\{\frac{133G}{9}, 2\|z^0 - z^*\|^2\right\},$$

and we shall prove the inequality by induction. For $k=0$, the inequality holds trivially

$$d_0 = \|z^0 - z^*\|^2 \leq \frac{Q}{2}.$$

Assuming the inequality holds for all index $1, ..., k$, we then have

$$\begin{aligned}
d_{k+1} &\leq \left(1 - \frac{55}{32(k+2)}\right) d_k + \frac{7}{32(k+2)} d_{k-1} + \frac{128G}{31(k+2)^2} + \frac{128\delta D}{31(k+2)\mu} \\
&\leq \left(1 - \frac{55}{32(k+2)}\right)\left(\frac{Q}{k+2} + \frac{256\delta D}{93\mu}\right) + \frac{7}{32(k+2)}\left(\frac{Q}{k+1} + \frac{256\delta D}{93\mu}\right) \\
&\quad + \frac{128G}{31(k+2)^2} + \frac{128\delta D}{31(k+2)\mu} \\
&\leq \left(1 - \frac{55}{32(k+2)}\right) \cdot \frac{Q}{k+2} + \frac{7}{32(k+2)} \cdot \frac{Q}{k+1} + \frac{128G}{31(k+2)^2} \\
&\quad + \frac{256\delta D}{93\mu} - \frac{440\delta D}{93(k+2)\mu} + \frac{56\delta D}{93(k+2)\mu} + \frac{128\delta D}{31(k+2)\mu} \\
&= \frac{Q}{k+2} - \frac{55Q}{32(k+2)^2} + \frac{7}{32(k+2)} \cdot \frac{Q}{k+1} + \frac{128G}{31(k+2)^2} + \frac{256\delta D}{93\mu} \\
&\leq \frac{Q}{k+3} + \frac{Q}{(k+2)^2} - \frac{55Q}{32(k+2)^2} + \frac{7}{32(k+2)} \cdot \frac{2Q}{k+2} + \frac{133G}{32(k+2)^2} + \frac{256\delta D}{93\mu} \\
&= \frac{Q}{k+3} - \frac{9Q}{32(k+2)^2} + \frac{133G}{32(k+2)^2} + \frac{256\delta D}{93\mu} \\
&= \frac{Q}{k+3} + \frac{256\delta D}{93\mu}.
\end{aligned}$$

Note that in the last inequality we used the identities $\frac{1}{k+1} \leq \frac{2}{k+2}$ and $\frac{128}{31} \leq \frac{133}{32}$. This completes the proof for the uniform $\mathcal{O}\left(\frac{1}{k}\right)$ convergence rate.

Chapter 5

An Accelerated Variance Reduced Extra-Point Approach to Finite-Sum VI Problems and Optimization

In this chapter, we develop stochastic variance reduced algorithms for solving a class of *finite-sum* monotone VI, where the operator consists of the sum of finitely many monotone VI mappings and the sum of finitely many monotone gradient mappings. We study the gradient complexities of the proposed algorithms under the settings when the sum of VI mappings is either strongly monotone or merely monotone. Furthermore, we consider the case when each of the VI mapping and gradient mapping is only accessible via noisy stochastic estimators and establish the sample gradient complexity. We demonstrate the application of the proposed algorithms for solving finite-sum convex optimization with finite-sum inequality constraints and develop a zeroth-order approach when only noisy and biased samples of objective/constraint function values are available.

5.1 Introduction

In this chapter, we consider an extended class of monotone VI (1.1.1) in the finite-sum form, where the operator consists of the sum of finitely many general vector mapping $H_i(x)$ and

the sum of finitely many gradient mapping $\nabla g_i(x)$:

$$F(x) = H(x) + \nabla g(x) := \sum_{i=1}^{m_1} H_i(x) + \sum_{i=1}^{m_2} \nabla g_i(x), \tag{5.1.1}$$

where each $H_i(\cdot)$ is Lipschitz continuous with constant $L_{h(i)}$ and $H(\cdot)$ is (strongly) monotone with constant $\mu\,(>) \geq 0$, and each $\nabla g_i(\cdot)$ is Lipschitz continuous with constant $L_{g(i)}$ and $\nabla g(\cdot)$ is monotone. The pioneering work considering such extended class of monotone VI (5.1.1) without the finite-sum structure (i.e. $m_1 = m_2 = 1$) is [10], where the authors propose a stochastic accelerated mirror-prox method with iteration (sample) complexity

$$\mathcal{O}\left(\sqrt{\frac{L_g}{\epsilon}} + \frac{L_h}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$$

for monotone $H(\cdot)$ and $\nabla g(\cdot)$. Note that the subscript i indicating the index in the finite-sum is omitted since $m_1 = m_2 = 1$. In addition, the authors of [10] consider the stochastic setting where both $H(\cdot)$ and $\nabla g(\cdot)$ can only be estimated via an unbiased oracle with bounded variance σ^2. In this chapter, we continue along this line of research on the extended class of monotone VI (5.1.1) with general m_1, m_2 and apply variance reduced techniques to establish accelerated gradient complexity bound. We assume each $H_i(\cdot)$ and $\nabla g_i(\cdot)$ can only be estimated via a stochastic oracle with bounded variance and bias and give the corresponding sample gradient complexity. We show that the proposed algorithms can be applied to solving finite-sum convex optimization with finite-sum inequality constraints [51] with an improved gradient complexity. Furthermore, the general stochastic setting in this chapter makes it possible to apply zeroth-order approach [46] to solve the aforementioned problem with our algorithm, when only biased samples of objective/constraint function values are accessible.

5.1.1 Algorithmic structure

The two proposed algorithms in this paper, SAVREP (5.2.7) and SAVREP-m (5.3.22), for solving strongly monotone and monotone problems respectively, have multiple-sequence updating structures. These "extra sequences" in our algorithms serve two main purposes:

- To accelerate the convergence to the (a) solution based on the composite VI mapping/gradient mapping structure in the problem class (5.1.1);
- To implement variance reduction when the VI mappings/gradient mappings have the finite-sum structure. That is, to reduce the estimation of each component mapping while achieving the first goal.

In order to achieve the first goal, it is critical that two different groups of sequences are used to accelerate the convergence in terms of the VI mappings and the gradient mappings respectively, because the two acceleration processes and the rates that are achieved differ in their respective problem classes. To achieve the second goal, it is again required to use two different groups of sequences to help control the number of either VI mappings/gradient mappings estimated during the iterations, which eventually guarantees the variance reduced gradient complexities in the respective problem classes.

With the above high-level concepts in mind, we can decompose the proposed algorithms into several major components:

1. **Acceleration of the VI mappings.**

   The most commonly used acceleration procedure for monotone VI problems is the extra-gradient-type method [41], which is also widely adopted in many VI algorithms that follow, e.g. [68, 62, 27, 28] and the references therein. This specific update scheme forms the basic structure in our algorithms:

$$\left\{\begin{array}{lcll} x^{k+0.5} & = & \arg\min\limits_{x\in\mathcal{Z}} & \gamma\langle H(x^k), x-x^k\rangle + \frac{1}{2}\|x-x^k\|^2, \\ x^{k+1} & = & \arg\min\limits_{x\in\mathcal{Z}} & \gamma\langle H(x^{k+0.5}), x-x^k\rangle + \frac{1}{2}\|x-x^k\|^2. \end{array}\right. \tag{5.1.2}$$

2. **Acceleration of the gradient mappings.**

   To realize the acceleration of the gradient mappings, Nesterov's acceleration in optimization [65, 66] is adopted in combination with the above extra-gradient method (5.1.2):

$$\left\{\begin{array}{lcll} y^k & = & (1-\alpha)v^k + \alpha x^k, & \\ x^{k+0.5} & = & \arg\min\limits_{x\in\mathcal{Z}} & \gamma\langle H(x^k) + \nabla g(y^k), x-x^k\rangle + \frac{1}{2}\|x-x^k\|^2, \\ x^{k+1} & = & \arg\min\limits_{x\in\mathcal{Z}} & \gamma\langle H(x^{k+0.5}) + \nabla g(y^k), x-x^k\rangle + \frac{1}{2}\|x-x^k\|^2, \\ v^{k+1} & = & (1-\alpha)v^k + \alpha x^{k+0.5}. & \end{array}\right. \tag{5.1.3}$$

   Note how the above update reduces to a regular extra-gradient method when $m_2 = 0$ (thus $\nabla g(\cdot) = 0$) and to a method similar to Nesterov's acceleration when $m_1 = 0$ (thus $H(\cdot) = 0$). Such concept of integration of acceleration from two different problem classes is first proposed in [10] and extended in [27].

3. **Variance reduction of the VI mappings.**

   The authors in [2] propose various variance reduced algorithms for finite-sum VI with $m_2 = 0$, including variance reduced extra-gradient method, whose main concept will be applied in our algorithms. The idea is to replace the sum of VI mappings $H(x^{k+0.5})$

with variance reduced gradient estimator $\hat{H}(x^{k+0.5})$ (formally defined in (5.2.12), Section 5.2.1) and use two other sequences, $\{\bar{x}^k\}$ and $\{w^k\}$, to help control the frequency of estimating the total mapping $H(\cdot)$. Incorporated into the combined acceleration scheme (5.1.3), we have:

$$\begin{cases} \bar{x}^k & = & (1-p_1)x^k + p_1 w^k \\ y^k & = & (1-\alpha)v^k + \alpha x^k \\ x^{k+0.5} & = & \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma\langle H(w^k) + \nabla g(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\ x^{k+1} & = & \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma\langle \hat{H}(x^{k+0.5}) + \nabla g(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\ v^{k+1} & = & (1-\alpha)v^k + \alpha x^{k+0.5} \\ w^{k+1} & = & \begin{cases} x^{k+1}, & \text{with prob. } p_1 \\ w^k, & \text{with prob. } 1-p_1. \end{cases} \end{cases}$$

4. **Variance reduction of the gradient mappings.**

In the case when m_2 is large, it is desirable to also apply variance reduction to the sum of gradient mappings $\nabla g(\cdot)$ in the update. The variance reduced gradient estimator $\tilde{\nabla} g(y^k)$ (formally defined in (5.2.13), Section 5.2.1) is used in the update instead of the total gradient estimator $\nabla g(y^k)$, together with an additional sequence $\{\bar{w}^k\}$ to control the frequency of estimating the total gradient estimator, as shown below:

$$\begin{cases} \bar{x}^k & = & (1-p_1)x^k + p_1 w^k \\ y^k & = & (1-\alpha-\beta)v^k + \alpha x^k + \beta \bar{w}^k \\ x^{k+0.5} & = & \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma\langle H(w^k) + \tilde{\nabla} g(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\ x^{k+1} & = & \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma\langle \hat{H}(x^{k+0.5}) + \tilde{\nabla} g(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\ v^{k+1} & = & (1-\alpha-\beta)v^k + \alpha x^{k+0.5} + \beta \bar{w}^k \\ w^{k+1} & = & \begin{cases} x^{k+1}, & \text{with prob. } p_1 \\ w^k, & \text{with prob. } 1-p_1 \end{cases} \\ \bar{w}^{k+1} & = & \begin{cases} v^{k+1}, & \text{with prob. } p_2 \\ \bar{w}^k, & \text{with prob. } 1-p_2. \end{cases} \end{cases} \tag{5.1.4}$$

Variance reduction in finite-sum optimization has been widely studied in the literature; see [3] and the references therein. When the problem is merely convex (or merely monotone VI), the last sequence $\{\bar{w}^k\}$ has to be modified such that it only updates after m_2 iterations:

$$\bar{w}^{k+1} = \begin{cases} \frac{1}{m_2}\sum_{i=k+2-m_2}^{k+1} v^i, & m_2|(k+1) \\ \bar{w}^k, & \text{otherwise.} \end{cases} \tag{5.1.5}$$

which forms a double-loop structure in the main update procedure (5.1.4). We present the detailed analysis in Section 5.3.

5. **Replacement of deterministic mappings with stochastic estimators.** Finally, we consider the more general stochastic setting in our algorithms, where each component in the finite-sum VI mapping $H(\cdot)$ and the finite-sum gradient mapping $\nabla g(\cdot)$ is allowed to be noisy stochastic estimators with bounded bias and variance. Denoted by $H'(\cdot)$ and $\nabla g'(\cdot)$, these estimators are in place of $H(\cdot)$ and $\nabla g(\cdot)$ in the update (5.1.4) to accommodate a different stochastic aspect from the finite-sum itself. Detailed definitions and assumptions are left to Section 5.2.1. The analysis that follows in Section 5.2.2 and 5.3.1 takes the stochastic errors into account, such that sample complexities can be readily derived from our results (see Proposition 5.2.4 and Corollary 5.3.5).

5.1.2 Structure of the analysis

Let us also briefly highlight the underlying high-level ideas of our analysis to facilitate understanding, while leaving detailed discussions to later sections.

The potential function. For a positive potential function P_k that depends on iteration k, the goal in the derivations is to establish $P_{k+1} \leq P_k$. Therefore, the form of the potential function becomes important in that it guides the logic behind the analysis. In our analysis, the potential function P_k is a combination of (the expectation of) $Q(x; x^*)$ and $\|x - x^*\|^2$, estimated at iteration k at different iterate points (since there are multiple sequences in the algorithms). In particular,

$$Q(x'; x) := \langle H(x), x' - x \rangle + g(x') - g(x),$$

where $Q(x'; x^*) \geq 0$ for any $x' \in \mathcal{X}$ (see (5.2.18)). In view of this specific form of $Q(x'; x)$, we derive upper bounds for $\langle H(x), x' - x \rangle$ and $g(x') - g(x)$ respectively, in order to establish the decrease in the potential function P_k.

The convergence in terms of VI mappings. Analyzing the sequences pertaining to the VI mappings is the key to deriving upper bounds for $\langle H(x), x' - x \rangle$. By our previous discussions, they are $\{x^k\}$, $\{x^{k+0.5}\}$, and $\{w^k\}$ ($\bar{x}^k$ are simply combination of x^k and w^k). Therefore, the first step of our analysis is to establish relations among these iterates; see Lemma 5.2.1.

The convergence in terms of gradient mappings. Identifying that the sequences $\{y^k\}$, $\{v^k\}$, and $\{\bar{w}^k\}$ are the ones that determine the convergence in terms of the gradient mappings, the next part of the analysis consists of deriving relations among these sequences, which leads to the upper bound for $g(x') - g(x)$; see Lemma 5.2.2.

(SAVREP-m) The convergence after an outer iteration. While in the strongly monotone problem, the final gradient complexity results can be readily derived after combining the results in Lemma 5.2.1 and 5.2.2 (see Theorem 5.2.3 and Proposition 5.2.4), the analysis in monotone problem needs to go an extra mile. As mentioned in the previous discussions, in the monotone problem the sequence $\{\bar{w}^k\}$ has to be updated in a double-loop structure (5.1.5), which is key to applying variance reduction for the gradient mappings. Therefore, we have to establish the relations of the (modified) potential function in terms of $\{\bar{w}^k\}$ after one outer iteration (which consists of m_2 inner iterations). This process includes identifying sophisticated relations among parameters and is discussed in detail in Section 5.3.1, with the results in Theorem 5.3.4.

Choices of parameters. To obtain the final gradient complexity results, we shall give specific parameter choices that satisfy the various conditions required in order to guarantee the decrease in the potential function. The specific parameters are summarized in Proposition 5.2.4 and Corollary 5.3.5.

5.2 Variance Reduced Scheme for Finite-Sum Strongly Monotone VI and Finite-Sum Monotone Gradients

In this section, we present our first variance reduced scheme for solving VI, where the operator $F(\cdot)$ takes the combined VI/gradient mapping form with finite-sum structure respectively (5.1.1). We assume the constraint set $\mathcal{Z}$ to be closed and convex, and the problem is summarized below:

$$\begin{cases} \text{find } x^* \text{ s.t.} \quad \langle F(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{Z}, \\ F(x) = H(x) + \nabla g(x) := \sum_{i=1}^{m_1} H_i(x) + \sum_{i=1}^{m_2} \nabla g_i(x). \end{cases} \tag{5.2.6}$$

We specifically consider the combined finite-sum operator $F(\cdot)$ being strongly monotone in this section, and we shall propose an alternative approach for $F(\cdot)$ being merely monotone in the next section. In particular, we assume $H(\cdot)$ to be strongly monotone with modulus $\mu_h > 0$, and $\nabla g(\cdot)$ to be monotone. Furthermore, we denote $H(\cdot) = \sum_{i=1}^{m_1} H_i(\cdot)$, where each $H_i(\cdot)$ is Lipschitz continuous with constant $L_{h(i)}$, and denote $g(x) = \sum_{i=1}^{m_2} g_i(x)$ (therefore $\nabla g(x) = \sum_{i=1}^{m_2} \nabla g_i(x)$) where each $\nabla g_i(x)$ is Lipschitz continuous with constant $L_{g(i)}$. Let us also define the sum of the Lipschitz constants $L_h := \sum_{i=1}^{m_1} L_{h(i)}$ and $L_g := \sum_{i=1}^{m_2} L_{g(i)}$.

Consider the following update for iteration count k:

$$
(5.2.7) \qquad \begin{cases}
\bar{x}^k & = & (1-p_1)x^k + p_1 w^k \\
y^k & = & (1-\alpha-\beta)v^k + \alpha x^k + \beta \bar{w}^k \\
x^{k+0.5} & = & \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma\langle H'(w^k) + \tilde{\nabla} g'(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\
x^{k+1} & = & \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma\langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\
v^{k+1} & = & (1-\alpha-\beta)v^k + \alpha x^{k+0.5} + \beta \bar{w}^k \\
w^{k+1} & = & \begin{cases} x^{k+1}, & \text{with prob. } p_1 \\ w^k, & \text{with prob. } 1-p_1 \end{cases} \\
\bar{w}^{k+1} & = & \begin{cases} v^{k+1}, & \text{with prob. } p_2 \\ \bar{w}^k, & \text{with prob. } 1-p_2. \end{cases}
\end{cases}
$$

Method (5.2.7) is a general *stochastic variance reduced* scheme for solving (5.2.6), and in the rest of this chapter we refer to it as **S**tochastic **A**ccelerated **V**ariance **R**educed **E**xtra **P**oint method (SAVREP). We shall make the following remarks. First, the variance reduction techniques are applied to *both* the general VI operator $H(\cdot)$ and the gradient mapping $\nabla g(\cdot)$, and the resulting update procedure will require using the variance reduced gradient estimator [3, 2], denote by $\hat{H}(\cdot)$ and $\tilde{\nabla} g(\cdot)$, respectively. Second, although the construction of the variance reduced gradient estimator $\hat{H}(\cdot)$ ($\tilde{\nabla} g(\cdot)$) involves sampling from the m_1 (m_2) individual operators $H_i(\cdot)$ ($\nabla g_i(\cdot)$) as we shall see later, we use the term "stochastic" to specifically refer to the fact that the update of SAVREP (5.2.7) only accesses the *noisy* estimations of the individual operators, denote by $H_i'(\cdot)$ and $\nabla g_i'(\cdot)$, respectively. This allows the application of zeroth-order approach [46] to problems when the gradients are unavailable and only the function values can be sampled, such as black-box optimization [72, 87, 17, 88] and saddle-point problem [99, 101, 54, 85, 56]. We shall exemplify such application in Section 5.4. As discussed in Section 5.1.1, the multiple-sequence structure is key to our algorithm, and the derivations of gradient complexity and sample complexity involving the analysis of each of these sequences are discussed in Section 5.2.2, following Section 5.2.1 where the detailed formulations of the (stochastic) variance reduced gradient estimators and the corresponding assumptions are presented.

5.2.1 Preliminaries

We first state the assumptions for the stochastic estimators $H_i'(\cdot)$ ($\nabla g_i'(\cdot)$) of the individual operators $H_i(\cdot)$ ($\nabla g_i(\cdot)$) for $i = 1, 2, ..., m_1$ (m_2). Denote $\mathbb{E}'[\cdot]$ as the expectation taken for

these samples and consider the following bias and variance upper bounds:

$$\left\|H_i(x) - \mathbb{E}'[H_i'(x)]\right\| \le \delta_h, \quad \mathbb{E}'\left[\left\|H_i'(x) - \mathbb{E}'\left[H_i'(x)\right]\right\|^2\right] \le \sigma_h^2, \tag{5.2.8}$$

$$\left\|\nabla g_i(x) - \mathbb{E}'\left[\nabla g_i'(x)\right]\right\| \le \delta_g, \quad \mathbb{E}'\left[\left\|\nabla g_i'(x) - \mathbb{E}'\left[\nabla g_i'(x)\right]\right\|^2\right] \le \sigma_g^2, \tag{5.2.9}$$

for some $\delta_h, \sigma_h^2, \delta_g, \sigma_g^2 \ge 0$. In other words, we assume the variance and the bias of these samples to be upper bounded by some non-negative constants (therefore they are not necessarily unbiased estimators). Denote $H'(x) := \sum_{i=1}^{m_1} H_i'(x)$ (respectively $\nabla g'(x) := \sum_{i=1}^{m_2} \nabla g_i'(x)$) as the sum of m_1 (respectively m_2) such independent stochastic oracles. The below bounds follow straightforwardly:

$$\left\|H(x) - \mathbb{E}'[H'(x)]\right\| \le m_1\delta_h, \quad \mathbb{E}'\left[\left\|H(x) - H'(x)\right\|^2\right] \le 2m_1\sigma_h^2 + 2m_1^2\delta_h^2, \tag{5.2.10}$$

$$\left\|\nabla g(x) - \mathbb{E}'[\nabla g'(x)]\right\| \le m_2\delta_g, \quad \mathbb{E}'\left[\left\|\nabla g(x) - \nabla g'(x)\right\|^2\right] \le 2m_2\sigma_g^2 + 2m_2^2\delta_g^2. \tag{5.2.11}$$

We next give explicit expressions for the (noiseless) variance reduced gradient estimators at the corresponding iterates given in (5.2.7):

$$\hat{H}(x^{k+0.5}) := H(w^k) + H_{\xi_k}(x^{k+0.5}) - H_{\xi_k}(w^k) \tag{5.2.12}$$

$$\tilde{\nabla} g(y^k) := \nabla g(\bar{w}^k) + \nabla g_{\zeta_k}(y^k) - \nabla g_{\zeta_k}(\bar{w}^k). \tag{5.2.13}$$

The above forms follow from the well-established variance reduction literature [3, 2], and the random variables $\xi\,(\zeta)$ take samples from the $m_1\,(m_2)$ individual operators $H_i(\cdot)\,(\nabla g_i(\cdot))$ with probability distribution taking respective Lipschitz constants $L_{h(i)}\,(L_{g(i)})$ into account. In particular, we have

$$\Pr\{\xi = i\} = \frac{L_{h(i)}}{L_h} := q_i,\ i = 1, 2, ..., m_1, \quad \Pr\{\zeta = i\} = \frac{L_{g(i)}}{L_g} := \pi_i,\ i = 1, 2, ..., m_2.$$

The stochastic oracles are then given by $H_\xi(\cdot) := \frac{1}{q_i} H_i(\cdot)$ and $\nabla g_\zeta(\cdot) = \frac{1}{\pi_i}\nabla g_i(\cdot)$.

However, note that in the update (5.2.7), only the noisy variance reduced gradient operators $\hat{H}'(x^{k+0.5})$ $(\tilde{\nabla} g'(y^k))$ are accessed, which are defined by:

$$\hat{H}'(x^{k+0.5}) := H'(w^k) + H'_{\xi_k}(x^{k+0.5}) - H'_{\xi_k}(w^k)$$
$$\tilde{\nabla} g'(y^k) := \nabla g'(\bar{w}^k) + \nabla g'_{\zeta_k}(y^k) - \nabla g'_{\zeta_k}(\bar{w}^k),$$

where $H'_\xi(\cdot) := \frac{1}{q_i} H_i'(\cdot)$ and $\nabla g'_\zeta(\cdot) = \frac{1}{\pi_i}\nabla g_i'(\cdot)$. To save the computational costs, we can reuse the noisy samples estimated at the same iterate within each iteration. For example, after sampling $H'(w^k)$ in the update of $x^{k+0.5}$, we could reuse the oracles for $H'_{\xi_k}(w^k)$ and $H'(w^k)$ in constructing $\hat{H}'(x^{k+0.5})$.

Finally, to simplify the notations in the following analysis, denote the expressions of conditional expectations taken for different random variables:

$$\mathbb{E}_{k_1}[\cdot] := \mathbb{E}_{\xi_k}[\cdot|x^k, w^k], \quad \mathbb{E}_{k_2}[\cdot] := \mathbb{E}_{\zeta_k}[\cdot|x^k, \bar{w}^k, v^k], \tag{5.2.14}$$

$$\mathbb{E}_{k_1+}[\cdot] := \mathbb{E}_{\xi_k}[\cdot|x^{k+1}, w^k], \quad \mathbb{E}_{k_2+}[\cdot] := \mathbb{E}_{\zeta_k}[\cdot|\bar{w}^k, v^{k+1}]. \tag{5.2.15}$$

5.2.2 Gradient complexity analysis

As highlighted in Section 5.1.2, the analysis first established one-iteration relation for the VI mappings, followed by one-iteration relation for gradient mappings, and then combining the two results in the decrease in the potential function. By selecting the parameters carefully, we derive the resulting gradient complexity for obtaining an ϵ-solution $\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \epsilon$, together with the corresponding stochastic errors. The lemma below summarizes the results from the first part of the analysis.

Lemma 5.2.1. *For the iterates generated by (5.2.7), define the following stochastic error terms:*

$$\varepsilon_x := \|H_\xi(x) - H'_\xi(x)\|, \quad \bar{\varepsilon}_x := \|H(x) - H'(x)\|.$$

Then, the following inequality holds for any $x \in \mathcal{Z}$ and $k = 0, 1, 2, ...$

$$\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma\langle H(x) + \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[(1 - p_1 - \frac{1}{2}\gamma\mu_h)\|x^k - x\|^2 + p_1\|w^k - x\|^2 - \|x^{k+1} - x\|^2\right] \\
& -\frac{1}{2}\left(p_1 - 2\gamma^2 L_h^2\right)\mathbb{E}_{k_1}\left[\|x^{k+0.5} - w^k\|^2\right] - \frac{1}{2}\left(1 - p_1 - \gamma\mu_h\right)\mathbb{E}_{k_1}\left[\|x^{k+0.5} - x^k\|^2\right] \\
& + \frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h} + \gamma^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}} + \varepsilon_{w^k})^2\right].
\end{aligned}$$

Proof. See Appendix 5.7.1. □

In Lemma 5.2.1, we define two stochastic error terms, ε_x and $\bar{\varepsilon}_x$, which are due to the noisy samples of $H_i(\cdot)$. We can first bound the squared stochastic error $\bar{\varepsilon}_x^2$ for the total operator $H(\cdot)$ with our assumption in (5.2.10):

$$\mathbb{E}'[\bar{\varepsilon}_x^2] \leq 2m_1\sigma_h^2 + 2m_1^2\delta_h^2.$$

On the other hand, the error ε_x involves a random variable ξ sampled from $i = 1, ..., m_1$. Since

$$\mathbb{E}_\xi\left[\varepsilon_x^2\right] = \sum_{i=1}^{m_1} q_i \cdot \frac{1}{q_i^2}\|H'_i(x) - H_i(x)\|^2 = \sum_{i=1}^{m_1}\frac{1}{q_i}\|H'_i(x) - H_i(x)\|^2,$$

we have

$$\begin{aligned}
\mathbb{E}'\left[\varepsilon_x^2\right] = \mathbb{E}'\left[\mathbb{E}_\xi\left[\varepsilon_x^2\right]\right] &\leq \sum_{i=1}^{m_1} \frac{2}{q_i}\mathbb{E}'\left[\left\|H_i'(x) - \mathbb{E}'\left[H_i'(x)\right]\right\|^2 + \left\|\mathbb{E}'\left[H_i'(x)\right] - H_i(x)\right\|^2\right] \\
&\leq 2(\sigma_h^2 + \delta_h^2)\cdot\sum_{i=1}^{m_1}\frac{1}{q_i} = 2L_h\cdot(\sigma_h^2+\delta_h^2)\cdot\sum_{i=1}^{m_1}\frac{1}{L_{h(i)}} := \tilde{\sigma}_h^2. \qquad (5.2.16)
\end{aligned}$$

Now we shall proceed to present the results in the second part of the analysis, summarized in the next lemma.

Lemma 5.2.2. *For the iterates generated by (5.2.7), define the following stochastic error terms:*

$$\rho_x := \|\nabla g_\zeta(x) - \nabla g_\zeta'(x)\|, \quad \bar{\rho}_x := \|\nabla g(x) - \nabla g'(x)\|.$$

With the condition $1-\alpha-\beta \geq 0$, the following inequality holds for any $x \in \mathcal{Z}$ and $k = 0, 1, 2, ...$

$$\begin{aligned}
&\mathbb{E}_{k_2}\left[g(v^{k+1}) - g(x)\right] \\
\leq\ &\mathbb{E}_{k_2}\left[(1-\alpha-\beta)\left(g(v^k) - g(x)\right) + \beta\left(g(\bar{w}^k) - g(x)\right)\right] \\
&+\mathbb{E}_{k_2}\left[\alpha\langle\tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] + \left(\frac{\alpha^2 L_g}{2} + \frac{\alpha^2 L_g}{2\beta} + \frac{\alpha\mu_h}{8}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5} - x^k\|^2\right] \\
&+\frac{\alpha\mu_h}{8}\mathbb{E}_{k_2}\left[\|x^k - x\|^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}_{\bar{w}^k}^2] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k} + \rho_{y^k})^2\right].
\end{aligned}$$

Proof. See Appendix 5.7.2. □

Similarly, we define the two stochastic error terms in Lemma 5.2.2, ρ_x and $\bar{\rho}_x$, which are due to the noisy samples of the gradient mapping $\nabla g_i(\cdot)$. The bound for $\bar{\rho}_x^2$ follows directly from (5.2.11):

$$\mathbb{E}'[\bar{\rho}_x^2] \leq 2m_2\sigma_g^2 + 2m_2^2\delta_g^2,$$

whereas the bound for ρ_x^2 can be derived as follows

$$\begin{aligned}
\mathbb{E}'\left[\mathbb{E}_\zeta[\rho_x^2]\right] &= \mathbb{E}'\left[\mathbb{E}_\zeta\left[\|\nabla g_\zeta(x) - \nabla g_\zeta'(x)\|^2\right]\right]\mathbb{E}'\left[\sum_{i=1}^{m_2}\pi_i\cdot\frac{1}{\pi_i^2}\|\nabla g_i(x) - \nabla g_i'(x)\|^2\right] \\
&\leq \sum_{i=1}^{m_2}\frac{2}{\pi_i}\mathbb{E}'\left[\left\|\nabla g_i(x) - \mathbb{E}'\left[\nabla g_i'(x)\right]\right\|^2 + \left\|\mathbb{E}'\left[\nabla g_i'(x)\right] - \nabla g_i'(x)\right\|^2\right] \\
&\leq 2(\sigma_g^2+\delta_g^2)\cdot\sum_{i=1}^{m_2}\frac{1}{\pi_i} = 2L_g\cdot(\sigma_g^2+\delta_g^2)\cdot\sum_{i=1}^{m_2}\frac{1}{L_{g(i)}} := \tilde{\sigma}_g^2. \qquad (5.2.17)
\end{aligned}$$

The last part of the analysis will combine the results from Lemma 5.2.1 and Lemma 5.2.2 and establish the overall per-iteration convergence in terms of a potential function. Let us first define the following function, which serves as an important component in our potential function:

$$Q(x';x) := \langle H(x), x'-x\rangle + g(x') - g(x).$$

In particular, we will use the function $Q(x';x^*)$ with x' being the iterates generated by SAVREP (5.2.7). The following properties show that $Q(x';x^*)$ is nonnegative for any $x' \in \mathcal{Z}$ and is upper-bounded in terms of x':

$$Q(x';x^*) = \langle H(x^*), x'-x^*\rangle + g(x') - g(x^*) \geq \langle H(x^*) + \nabla g(x^*), x'-x^*\rangle \geq 0. \tag{5.2.18}$$

and

$$\begin{aligned} Q(x';x^*) &= \langle H(x^*), x'-x^*\rangle + g(x') - g(x^*) \\ &\leq \langle H(x'), x'-x^*\rangle - \mu_h\|x'-x^*\|^2 + g(x') - g(x^*) \\ &\leq \langle H(x') + \nabla g(x'), x'-x^*\rangle - \mu_h\|x'-x^*\|^2 \leq \frac{1}{4\mu_h}\left\|H(x') + \nabla g(x')\right\|^2. \end{aligned}$$

Now we are ready to show the per-iteration convergence for (5.2.7):

Theorem 5.2.3. *For the iterates generated by (5.2.7), define the following constants:*

$$\Delta_h := \frac{\alpha}{\mu_h}(m_1\sigma_h^2 + m_1^2\delta_h^2) + 2\alpha\gamma\tilde{\sigma}_h^2, \quad \Delta_g := \frac{16\alpha}{\mu_h}(m_2\sigma_g^2 + m_2^2\delta_g^2) + \frac{16\alpha}{\mu_h}\tilde{\sigma}_g^2.$$

Then, the following inequality holds for $k = 0, 1, 2, ...$

$$\begin{aligned} &\mathbb{E}\left[(1-\phi p_2)Q(v^{k+1};x^*) + \phi Q(\bar{w}^{k+1};x^*)\right] \\ &+\frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^{k+1}-x^*\|^2 + \|w^{k+1}-x^*\|^2\right] \\ \leq\ &\mathbb{E}\left[(1-\alpha-\beta)Q(v^k;x^*) + (\beta+\phi(1-p_2))Q(\bar{w}^k;x^*)\right] \\ &+\left(1-\frac{\gamma\mu_h}{12}\right)\frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^k-x^*\|^2 + \|w^k-x^*\|^2\right] + \Delta_h + \Delta_g. \end{aligned} \tag{5.2.19}$$

Proof. See Appendix 5.7.3. □

Theorem 5.2.3 establishes the relation for the iterates generated by (5.2.7), with additional stochastic errors Δ_h, Δ_g due to the noisy samples taken for $H_i(\cdot)$ and $\nabla g_i(\cdot)$. To further

derive the gradient complexity and the overall stochastic errors, we are left with specifying the parameters $\alpha, \beta, \gamma, \phi, p_1, p_2$. Note that in deriving (5.2.19), we have imposed the constraints (5.7.59) on some of the parameters (see Appendix 5.7.3), together with the condition $1 - \alpha - \beta \geq 0$ in Lemma 5.2.2, which should be honored during the parameter selection process. We summarize the gradient complexity results in the next proposition:

Proposition 5.2.4. *In view of Theorem 5.2.3, by specifying the following parameters:*

$$\gamma = \frac{1}{4}\min\left(\frac{\sqrt{p_1}}{L_h}, \sqrt{\frac{p_2}{L_g\mu_h}}, \frac{p_1}{\mu_h}\right), \quad \alpha = \frac{1}{12}\min\left(\sqrt{\frac{\mu_h}{L_g p_2}}, 1\right), \quad \beta = \frac{1}{2},$$

and

$$\phi = \frac{(1+\alpha)m_2}{2}, \quad p_1 = \frac{1}{m_1}, \quad p_2 = \frac{1}{m_2},$$

the gradient complexity for reducing the deterministic errors to some $\epsilon > 0$ is

$$\mathcal{O}\left(\left(m_1 + m_2 + \sqrt{\frac{L_g m_2}{\mu_h}} + \frac{L_h\sqrt{m_1}}{\mu_h}\right)\log\frac{d_0}{\epsilon}\right), \tag{5.2.20}$$

where

$$d_0 := \frac{\gamma}{\alpha\mu_h}\left\|H(x^0) + \nabla g(x^0)\right\|^2 + 2\|x^0 - x^*\|^2.$$

In addition, the overall stochastic error after reducing the deterministic error to ϵ is of the order

$$\mathcal{O}\left(\left(m_1 + m_2 + \sqrt{\frac{L_g m_2}{\mu_h}} + \frac{L_h\sqrt{m_1}}{\mu_h}\right)\cdot\frac{\gamma}{\alpha}\cdot(\Delta_h + \Delta_g)\right). \tag{5.2.21}$$

Proof. See Appendix 5.7.4. □

A few remarks are in order to interpret the results in Proposition 5.2.4.

Remark 5.2.5. *Under the noiseless case where $H_i(\cdot)$ and $\nabla g_i(\cdot)$ can be computed exactly ($\delta_h = \sigma_h = \delta_g = \sigma_g = 0$), (5.2.20) gives the iteration complexity before reaching either $\|x^k - x^*\|^2 \leq \epsilon$ or $\|w^k - w^*\|^2 \leq \epsilon$. Since in each iteration the full operator $H(\cdot)$ ($\nabla g(\cdot)$) is estimated at $w^k/\bar{w}^k$, which in expectation only updates every m_1 (m_2) iterations, the expected cost for estimating an individual operator $H_i(\cdot)$ ($\nabla g_i(\cdot)$) is constant. Therefore, (5.2.20) is also the gradient complexity for obtaining the ϵ-solution.*

For a general strongly monotone VI, [2] has established the $\mathcal{O}\left(\left(m_1 + \frac{L_h\sqrt{m_1}}{\mu_h}\right)\log\frac{1}{\epsilon}\right)$ gradient complexity, while for strongly convex optimization [3, 109] the gradient complexity $\mathcal{O}\left(\left(m_2 + \sqrt{\frac{L_g m_2}{\mu_g}}\right)\log\frac{1}{\epsilon}\right)$ has been established. While the former gradient complexity has not been shown tight for VI, Proposition 5.2.4 implies that it is indeed possible to improve upon the previous results and reflect the accelerated complexity from optimization, when the VI is of the specific form (5.2.6).

Remark 5.2.6. *Under the noisy case when the operators can only be estimated inexactly, the stochastic error $\Delta_h + \Delta_g$ will be carried throughout the iterations. Provided that the total number of iterations is in the order (5.2.20), the overall error is given by $\epsilon + \Delta_T$, where we refer to ϵ as the "deterministic error". The order of the overall "stochastic error" Δ_T, is then given by (5.2.21). Through standard techniques such as increasing the sample size for $H_i'(\cdot)$ ($\nabla g_i'(\cdot)$), the overall stochastic error Δ_T can be further reduced to $\mathcal{O}(\epsilon)$.*

5.3 Variance Reduced Scheme for Finite-Sum Monotone VI and Finite-Sum Monotone Gradients

In this section, we develop a new algorithm for the same finite-sum monotone VI in the form (5.2.6), but now we only assume $H(\cdot)$ to be monotone instead of strongly monotone, i.e. $\mu_h = 0$ (the monotone assumption for $\nabla g(\cdot)$ remains). The loss of strong monotonicity assumption therefore requires a different design of update procedure and analysis from the previous section, as we shall present shortly later. Same as in the previous section, we define $H(\cdot) = \sum_{i=1}^{m_1} H_i(\cdot)$ where each $H_i(\cdot)$ is Lipschitz continuous with constant $L_{h(i)}$, and $g(x) = \sum_{i=1}^{m_2} g_i(x)$ is sum of Lipschitz continuous gradient mappings, each with Lipschitz constant $L_{g(i)}$. The rest of the setups in Section 5.2.1 also apply, and we shall only supplement with some specific changes in the analysis that follows.

Consider the following update for iteration count k:

$$
\text{(5.3.22)} \quad \begin{cases}
\bar{x}^k &= (1-p_1)x^k + p_1 w^k \\
y^k &= (1-\alpha_k-\beta_k)v^k + \alpha_k x^k + \beta_k \bar{w}^k \\
x^{k+0.5} &= \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma_k\langle H'(w^k) + \tilde{\nabla} g'(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\
x^{k+1} &= \arg\min\limits_{x\in\mathcal{Z}} \quad \gamma_k\langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - \bar{x}^k\rangle + \frac{1}{2}\|x-\bar{x}^k\|^2 \\
v^{k+1} &= (1-\alpha_k-\beta_k)v^k + \alpha_k x^{k+0.5} + \beta_k \bar{w}^k \\
w^{k+1} &= \begin{cases} x^{k+1}, & \text{with prob. } p_1 \\ w^k, & \text{with prob. } 1-p_1 \end{cases} \\
\bar{w}^{k+1} &= \begin{cases} \frac{1}{m_2}\sum_{i=k+2-m_2}^{k+1} v^i, & m_2|(k+1) \\ \bar{w}^k, & \text{otherwise.} \end{cases}
\end{cases}
$$

There are two main differences between the update (5.3.22) presented above and the update (5.2.7) in the previous section. First, while (5.2.7) simply updates $\bar{w}^k$ with probability $p_2 = \frac{1}{m_2}$ in each iteration, (5.3.22) has a double-loop structure, which updates $\bar{w}^k$ once

every m_2 iterations. In other words, the full gradient $\nabla g(\bar{w}^k)$ is only estimated at the beginning of each outer-loop, and such gradient is used to obtain the variance reduced gradient $\tilde{\nabla} g(y^k)$ within each inner-loop. Second, instead of using constant parameters as in (5.2.7), the update (5.3.22) uses parameters $\alpha_k, \beta_k, \gamma_k$ that depend on iteration number k. These changes turn out to be critical in the monotone VI setting. We shall refer to the update (5.3.22) as SAVREP-m (SAVREP for monotone VI) in the rest of the paper.

5.3.1 Gradient complexity analysis

In order to establish a theoretical guarantee for the gradient complexity, we make two additional assumptions compared to the analysis of SAVREP. In particular, we assume that the stochastic estimators $H_i'(x)$ and $\nabla g_i'(x)$ are unbiased, and the constraint set $\mathcal{Z}$ is bounded, as summarized below.

Assumption 5.3.1. *The stochastic estimators $H_i'(x)$ and $\nabla g_i'(x)$ are both unbiased, i.e. $\delta_h = \delta_g = 0$ in (5.2.8)-(5.2.9).*

Assumption 5.3.2. *The diameter of the constraint set $\mathcal{Z}$ is $\Omega_{\mathcal{Z}}$, i.e.*

$$\sup_{x,y\in\mathcal{Z}} \|x - y\| = \Omega_{\mathcal{Z}}. \tag{5.3.23}$$

The gradient complexity analysis of SAVREP-m (5.3.22) consists of two major steps. The first step is to derive the per-iteration relation among iterates and establish a result similar to Theorem 5.2.3. In the first step, we only consider the iterations from k to $k+1$, which is within a single inner-loop in the update (5.3.22) with $\bar{w}^k$ remaining unchanged. In the second step, we derive the relation among iterates after one outer-loop, where the iterations proceed from sm_2 to $(s+1)m_2$. This step specifically establishes an inequality relating $\bar{w}^{(s+1)m_2}$ and $\bar{w}^{sm_2}$, which eventually guarantees the convergence of the iterate $\bar{w}^k$ as long as the parameters are chosen to satisfy certain conditions.

The results derived from the first step is presented in the next lemma:

Lemma 5.3.3. *For the iterates generated by (5.3.22), assume the following condition holds for all $k \geq 0$:*

$$\begin{cases} p_1 - 2\gamma_k^2 L_h^2 \geq 0, \\ q - p_1 - \alpha_k\gamma_k L_g - \frac{\alpha_k\gamma_k L_g}{\beta_k} \geq 0, \\ 1 - \alpha_k - \beta_k \geq 0 \end{cases} \tag{5.3.24}$$

where $0 < q < 1$ is a constant independent of the problem and algorithm parameters. Then

we have:

(5.3.25)

$$\begin{aligned}
&\mathbb{E}\left[Q(v^{k+1};x)\right] + \frac{\alpha_k}{2\gamma_k}\mathbb{E}\left[(1-p_1)\|x^{k+1}-x\|^2 + \|w^{k+1}-x\|^2\right] \\
\leq\ & \mathbb{E}\left[(1-\alpha_k-\beta_k)Q(v^k;x) + \beta_k Q(\bar{w}^k;x)\right] + \frac{\alpha_k}{2\gamma_k}\mathbb{E}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] + \alpha_k\gamma_k\Delta
\end{aligned}$$

where Δ is the stochastic error defined as:

$$\Delta = 2\tilde{\sigma}_h^2 + \frac{1}{(1-q)}(2m_2\sigma_g^2 + 4\tilde{\sigma}_g^2) = O\left(\sigma_h^2 L_h \sum_{i=1}^{m_1}\frac{1}{L_{h(i)}} + \sigma_g^2 L_g \sum_{i=1}^{m_2}\frac{1}{L_{g(i)}}\right).$$

Proof. See Appendix 5.7.5. □

Note that while Lemma 5.3.3 establishes the relation of iterates between iteration k and $k+1$, $\bar{w}^k$ remains unchanged (unless $m_2|k+1$). Since $\bar{w}^k$ plays the central role in the convergence under the monotone case, we have to extend the result in (5.3.25) to iterations between sm_2 and $(s+1)m_2$, where s denotes the number of outer-loops (or *epochs*). In particular, we assume that the parameters $\alpha_k, \beta_k, \gamma_k$ are also unchanged within each interval of updating $\bar{w}^k$, i.e. $\alpha_{sm_2} = \alpha_{sm_2+1} = \cdots = \alpha_{(s+1)m_2-1}$, $\beta_{sm_2} = \beta_{sm_2+1} = \cdots = \beta_{(s+1)m_2-1}$, and $\gamma_{sm_2} = \gamma_{sm_2+1} = \cdots = \gamma_{(s+1)m_2-1}$. Then, by summing up inequality (5.3.25) from $k = sm_2$ to $k = (s+1)m_2 - 1$, we get

$$\begin{aligned}
&\mathbb{E}\left[Q(v^{(s+1)m_2};x) + (\alpha_{sm_2}+\beta_{sm_2})\sum_{k=sm_2+1}^{(s+1)m_2-1} Q(v^k;x)\right] \\
&+\frac{\alpha_{sm_2}}{2\gamma_{sm_2}}\mathbb{E}\left[(1-p_1)\|x^{(s+1)m_2}-x\|^2 + \|w^{(s+1)m_2}-x\|^2\right] \\
\leq\ & (1-\alpha_{sm_2}-\beta_{sm_2})\mathbb{E}[Q(v^{sm_2};x)] + \beta_{sm_2}m_2\mathbb{E}[Q(\bar{w}^{sm_2};x)] \\
&+\frac{\alpha_{sm_2}}{2\gamma_{sm_2}}\mathbb{E}\left[(1-p_1)\|x^{sm_2}-x\|^2 + \|w^{sm_2}-x\|^2\right] + m_2\alpha_{sm_2}\gamma_{sm_2}\Delta. \qquad (5.3.26)
\end{aligned}$$

Since $Q(x';x) := \langle H(x), x'-x\rangle + g(x') - g(x)$ and g is convex, $Q(\cdot;x)$ is convex. By using the definition $\bar{w}^{sm_2} = \frac{1}{m_2}\sum_{i=(s-1)m_2+1}^{sm_2} v^i$ and the convexity of $Q(\cdot;x)$, we have

$\sum_{k=(s-1)m_2+1}^{sm_2} Q(v^k;x) \geq m_2 Q(\bar{w}^{sm_2};x)$. Then,

$$\begin{aligned}
&\mathbb{E}\left[Q(v^{(s+1)m_2};x) + (\alpha_{sm_2}+\beta_{sm_2})\sum_{k=sm_2+1}^{(s+1)m_2-1} Q(v^k;x)\right] \\
&+\frac{\alpha_{sm_2}}{2\gamma_{sm_2}}\mathbb{E}[(1-p_1)\|x^{(s+1)m_2}-x\|^2 + \|w^{(s+1)m_2}-x\|^2] \\
\leq\ & (1-\alpha_{sm_2})\mathbb{E}[Q(v^{sm_2};x)] + \beta_{sm_2}\mathbb{E}\left[\sum_{k=(s-1)m_2+1}^{sm_2-1} Q(v^k;x)\right] \\
&+\frac{\alpha_{sm_2}}{2\gamma_{sm_2}}\mathbb{E}[(1-p_1)\|x^{sm_2}-x\|^2 + \|w^{sm_2}-x\|^2] + m_2\alpha_{sm_2}\gamma_{sm_2}\Delta. \qquad (5.3.27)
\end{aligned}$$

Let us define

$$\Gamma_s = \begin{cases} 1, & \text{when } s=0 \\ (1-\alpha_{(s-1)m_2})\Gamma_{s-1}, & \text{when } s>0 \end{cases}$$

and

$$V(x',w;x) = (1-p_1)\|x'-x\|^2 + \|w-x\|^2.$$

Then, by dividing Γ_{s+1} on both sides of 5.3.27, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\Gamma_{s+1}}Q(v^{(s+1)m_2};x) + \frac{\alpha_{sm_2}+\beta_{sm_2}}{\Gamma_{s+1}}\sum_{k=sm_2+1}^{(s+1)m_2-1} Q(v^k;x)\right] \\
\leq\ & \frac{1}{\Gamma_s}\mathbb{E}[Q(v^{sm_2};x)] + \frac{\beta_{sm_2}}{\Gamma_{s+1}}\mathbb{E}\left[\sum_{k=(s-1)m_2+1}^{sm_2-1} Q(v^k;x)\right] \\
&+\frac{\alpha_{sm_2}}{2\gamma_{sm_2}\Gamma_{s+1}}\mathbb{E}\left[V(x^{sm_2},w^{sm_2};x) - V(x^{(s+1)m_2},w^{(s+1)m_2};x)\right] + \frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta. \qquad (5.3.28)
\end{aligned}$$

Since for any solution x^* we have $Q(x;x^*) \geq 0$ for all $x \in \mathcal{Z}$ (c.f. (5.2.18)), by taking $x = x^*$ in (5.3.28) with the condition on the parameters:

$$\frac{\beta_{sm_2}}{\Gamma_{s+1}} \leq \frac{\alpha_{(s-1)m_2}+\beta_{(s-1)m_2}}{\Gamma_s}, \qquad (5.3.29)$$

we can rewrite (5.3.28) into:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\Gamma_{s+1}}Q(v^{(s+1)m_2};x^*) + \frac{\alpha_{sm_2}+\beta_{sm_2}}{\Gamma_{s+1}}\sum_{k=sm_2+1}^{(s+1)m_2-1} Q(v^k;x^*)\right] \\
\leq\ & \frac{1}{\Gamma_s}\mathbb{E}[Q(v^{sm_2};x^*)] + \frac{\alpha_{(s-1)m_2}+\beta_{(s-1)m_2}}{\Gamma_s}\mathbb{E}\left[\sum_{k=(s-1)m_2+1}^{sm_2-1} Q(v^k;x^*)\right] \\
&+\frac{\alpha_{sm_2}}{2\gamma_{sm_2}\Gamma_{s+1}}\mathbb{E}\left[V(x^{sm_2},w^{sm_2};x^*) - V(x^{(s+1)m_2},w^{(s+1)m_2};x^*)\right] + \frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta. \qquad (5.3.30)
\end{aligned}$$

Now, assume the next condition to hold for $s = 1, ..., S-1$:

$$B_s := \frac{\alpha_{sm_2}}{2\gamma_{sm_2}\Gamma_{s+1}}, \quad B_{s-1} \le B_s. \tag{5.3.31}$$

Then we can obtain the next inequalities by summing up (5.3.30) for $s = 1, ..., S-1$:

$$\begin{aligned}
& \frac{m_2\beta_{(S-1)m_2}}{\Gamma_S}\mathbb{E}\left[Q(\bar{w}^{Sm_2}; x^*)\right] \\
\le\ & \mathbb{E}\left[\frac{1}{\Gamma_S}Q(v^{Sm_2}; x^*) + \frac{\alpha_{(S-1)m_2} + \beta_{(S-1)m_2}}{\Gamma_S}\sum_{k=(S-1)m_2+1}^{Sm_2-1} Q(v^k; x^*)\right] \\
\le\ & \frac{1}{\Gamma_1}\mathbb{E}[Q(v^{m_2}; x^*)] + \frac{\alpha_0 + \beta_0}{\Gamma_1}\mathbb{E}\left[\sum_{k=1}^{m_2-1} Q(v^k; x^*)\right] \\
& + \sum_{s=1}^{S-1} B_s\mathbb{E}\left[V(x^{sm_2}, w^{sm_2}; x^*) - V(x^{(s+1)m_2}, w^{(s+1)m_2}; x^*)\right] + \sum_{s=1}^{S-1}\frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta \\
\le\ & \frac{(1-\alpha_0-\beta_0)}{\Gamma_1}Q(v^0; x^*) + \frac{\beta_0 m_2}{\Gamma_1}Q(\bar{w}^0; x^*) \\
& + \sum_{s=0}^{S-1} B_s\mathbb{E}\left[V(x^{sm_2}, w^{sm_2}; x^*) - V(x^{(s+1)m_2}, w^{(s+1)m_2}; x^*)\right] + \sum_{s=0}^{S-1}\frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta \\
\le\ & \frac{(1-\alpha_0+(m_2-1)\beta_0)}{\Gamma_1}Q(w^0; x^*) + B_0\mathbb{E}\left[V(x^0, w^0; x^*)\right] \\
& + \sum_{s=1}^{S-1}(B_s - B_{s-1})\mathbb{E}\left[V(x^{sm_2}, w^{sm_2}; x^*)\right] + \sum_{s=0}^{S-1}\frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta \\
\overset{(5.3.23),(5.3.31)}{\le}\ & \frac{(1-\alpha_0+(m_2-1)\beta_0)}{\Gamma_1}Q(w^0; x^*) + B_0\Omega_Z^2 + \sum_{s=1}^{S-1}(B_s - B_{s-1})\Omega_Z^2 \\
& + \sum_{s=0}^{S-1}\frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta \\
\le\ & \frac{(1-\alpha_0+(m_2-1)\beta_0)}{\Gamma_1}Q(w^0; x^*) + \frac{\alpha_{(S-1)m_2}}{2\gamma_{(S-1)m_2}\Gamma_S}\Omega_Z^2 + \sum_{s=0}^{S-1}\frac{m_2\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}}\Delta,
\end{aligned}$$

where in the third inequality we apply (5.3.26) with $s = 0$ and $x = x^*$, and in the forth inequality we simply remove the nonpositive term $-B_{S-1}\mathbb{E}\left[V(x^{Sm_2}, w^{Sm_2}; x^*)\right]$, together with the definition $v^0 := \bar{w}^0 := w^0 = x^0$.

We summarize the above results together with the required conditions on the parameters (5.3.24), (5.3.29), (5.3.31) in the next theorem:

Theorem 5.3.4. *Suppose the following conditions hold for $k \ge 0$ and $s = 1, ..., S-1$:*

(5.3.32)

$$\begin{cases} p_1 - 2\gamma_k^2 L_h^2 & \ge\ 0 \\ q - p_1 - \alpha_k\gamma_k L_g - \frac{\alpha_k\gamma_k L_g}{\beta_k} & \ge\ 0 \\ 1 - \alpha_k - \beta_k & \ge\ 0 \end{cases}, \quad \begin{cases} \frac{\alpha_{(s-1)m_2}}{\gamma_{(s-1)m_2}\Gamma_s} & \le\ \frac{\alpha_{sm_2}}{\gamma_{sm_2}\Gamma_{s+1}} \\ \frac{\beta_{sm_2}}{1-\alpha_{sm_2}} & \le\ \alpha_{(s-1)m_2} + \beta_{(s-1)m_2} \end{cases}$$

where $0 < q < 1$ is a constant, and α_k, β_k, γ_k are constants within each interval of updating $\bar{w}$, i.e. $\alpha_{sm_2} = \alpha_{sm_2+1} = \cdots = \alpha_{(s+1)m_2-1}$, $\beta_{sm_2} = \beta_{sm_2+1} = \cdots = \beta_{(s+1)m_2-1}$, and $\gamma_{sm_2} = \gamma_{sm_2+1} = \cdots = \gamma_{(s+1)m_2-1}$. Then,

$$\begin{aligned} & \mathbb{E}\left[Q(\bar{w}^{Sm_2}; x^*)\right] \\ \leq \; & \frac{1}{m_2\beta_{(S-1)m_2}} \frac{(1-\alpha_0+(m_2-1)\beta_0)\Gamma_S}{\Gamma_1} Q(w^0; x^*) + \frac{\alpha_{(S-1)m_2}}{2m_2\gamma_{(S-1)m_2}\beta_{(S-1)m_2}}\Omega_Z^2 \\ & + \frac{\Gamma_S}{\beta_{(S-1)m_2}} \sum_{s=0}^{S-1} \frac{\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}} \Delta \end{aligned}$$

for any $x^* \in \mathcal{Z}^*$, where $\mathcal{Z}^* \subseteq \mathcal{Z}$ is the solution set and $\Delta = O\left(\sigma_h^2 L_h \sum_{i=1}^{m_1} \frac{1}{L_{h(i)}} + \sigma_g^2 L_g \sum_{i=1}^{m_2} \frac{1}{L_{g(i)}}\right)$.

We shall specify a set of parameters that satisfy the conditions in (5.3.32) and the corresponding gradient complexities in the next corollary.

Corollary 5.3.5. *If we choose*

$$\begin{aligned} & q = \frac{3}{4}, \quad p_1 = \frac{1}{m_1} \leq \frac{1}{2}, \quad \alpha_k = \frac{2}{s+4} \leq \frac{1}{2}, \quad \beta_k = \frac{1}{2}, \\ & \gamma_k = \frac{s+3}{24(L_g + (s+1)L_h\sqrt{m_1}) + (s+1)\sqrt{(s+1)\Delta m_2}/\Omega_{\mathcal{Z}}}, \end{aligned}$$

where $s = \left\lfloor \frac{k}{m_2} \right\rfloor$, then when $m_2|k$,

$$\begin{aligned} \mathbb{E}[Q(\bar{w}^k, x^*)] \; & \leq \; \frac{24}{S^2} Q(w^0, x^*) + \frac{48}{m_2 S^2} L_g \Omega_{\mathcal{Z}}^2 + \frac{48}{m_2 S} L_h \sqrt{m_1} \Omega_{\mathcal{Z}}^2 + \frac{26\Omega_{\mathcal{Z}}\sqrt{\Delta}}{\sqrt{Sm_2}} \\ & = \; \frac{24m_2^2}{k^2} Q(w^0, x^*) + \frac{48m_2}{k^2} L_g \Omega_{\mathcal{Z}}^2 + \frac{48}{k} L_h \sqrt{m_1} \Omega_{\mathcal{Z}}^2 + \frac{26\Omega_{\mathcal{Z}}\sqrt{\Delta}}{\sqrt{k}} \end{aligned} \tag{5.3.33}$$

where $S = k/m_2$. The gradient complexity for reducing $\mathbb{E}[Q(\bar{w}^k, x^*)]$ to some $\epsilon > 0$ is given by

$$\mathcal{O}\left(\sqrt{\frac{Q(w^0, x^*)}{\epsilon}} m_2 + \sqrt{\frac{L_g m_2}{\epsilon}} \Omega_{\mathcal{Z}} + \frac{L_h\sqrt{m_1}\Omega_{\mathcal{Z}}^2}{\epsilon} + \frac{\Delta\Omega_{\mathcal{Z}}^2}{\epsilon^2}\right). \tag{5.3.34}$$

Proof. We first verify the conditions (5.3.32) are satisfied by the specific choices of the parameters. Note that $\Gamma_s = \frac{6}{(s+2)(s+3)}$, and the following inequalities hold:

$$\gamma_k^2 L_h^2 \leq \left(\frac{s+3}{24(s+1)\sqrt{m_1}}\right)^2 \leq \frac{1}{2m_1} = \frac{p_1}{2},$$

$$p_1 + \alpha_k\gamma_k L_g + \frac{\alpha_k\gamma_k L_g}{\beta_k} = p_1 + 3\alpha_k\gamma_k L_g \le \frac{1}{2} + \frac{6}{s+4}\cdot\frac{s+3}{24} \le \frac{3}{4},$$

$$\frac{\alpha_{sm_2}}{\gamma_{sm_2}\Gamma_{s+1}} = 8(L_g + (s+1)L_h\sqrt{m_1}) + \sqrt{(s+1)\Delta m_2}/\Omega_{\mathcal{Z}}$$

which is non-decreasing in $s = 0, 1, ..., S-1$, and

$$\frac{s+4}{2(s+2)} = \frac{\beta_{sm_2}}{1-\alpha_{sm_2}} \le \alpha_{(s-1)m_2} + \beta_{(s-1)m_2} = \frac{s+7}{2(s+3)}.$$

Therefore, the conditions in (5.3.32) are indeed satisfied.

The convergence rate (5.3.33) can be derived by noticing the next few inequalities:

$$\frac{1}{m_2\beta_{(S-1)m_2}}\frac{(1-\alpha_0+(m_2-1)\beta_0)\Gamma_S}{\Gamma_1} \le 4\Gamma_S \le \frac{24}{S^2},$$

$$\begin{aligned}\frac{\alpha_{(S-1)m_2}}{2m_2\gamma_{(S-1)m_2}\beta_{(S-1)m_2}} &\le \frac{2}{m_2S^2}\left(24(L_g + SL_h\sqrt{m_1}) + S\sqrt{S\Delta m_2}/\Omega_{\mathcal{Z}}\right)\\ &\le \frac{48}{m_2S^2}L_g + \frac{48}{m_2S}L_h\sqrt{m_1} + \frac{2\sqrt{\Delta}}{\Omega_{\mathcal{Z}}\sqrt{Sm_2}},\end{aligned}$$

and

$$\frac{\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}} \le \frac{s+3}{3}\gamma_{sm_2} \le \frac{(s+3)^2}{3(s+1)\sqrt{(s+1)\Delta m_2}/\Omega_{\mathcal{Z}}} \le 3\sqrt{s+1}\frac{\Omega_{\mathcal{Z}}}{\sqrt{\Delta m_2}},$$

which results in the following bound since $\sum_{s=0}^{S-1}\sqrt{s+1} \le \int_{s=0}^{S+1}\sqrt{s}ds = \frac{2}{3}(S+1)^{3/2}$:

$$\frac{\Gamma_S}{\beta_{(S-1)m_2}}\sum_{s=0}^{S-1}\frac{\alpha_{sm_2}\gamma_{sm_2}}{\Gamma_{s+1}} \le 24\frac{\Omega_{\mathcal{Z}}}{\sqrt{S\Delta m_2}}.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[Q(\bar{w}^k, x^*)] &\le \frac{24}{S^2}Q(w^0, x^*) + \frac{48}{m_2S^2}L_g\Omega_{\mathcal{Z}}^2 + \frac{48}{m_2S}L_h\sqrt{m_1}\Omega_{\mathcal{Z}}^2 + \frac{26\Omega_{\mathcal{Z}}\sqrt{\Delta}}{\sqrt{Sm_2}}\\ &= \frac{24m_2^2}{k^2}Q(w^0, x^*) + \frac{48m_2}{k^2}L_g\Omega_{\mathcal{Z}}^2 + \frac{48}{k}L_h\sqrt{m_1}\Omega_{\mathcal{Z}}^2 + \frac{26\Omega_{\mathcal{Z}}\sqrt{\Delta}}{\sqrt{k}}.\end{aligned}$$

□

Remark 5.3.6. *In case of $H(x) = 0$ and $\Delta = 0$, the only difference between the above complexity and the counterpart of Katyusha [3] is we replace $\|w^0 - x^*\|$ with $\Omega_{\mathcal{Z}}$. In addition,*

when $H(x) \neq 0$, the complexity improves the result in [2] in terms of L_g. In the case of $g(x) = 0$, the complexity matches the results in [2], with the gap between the current lower bound [100] remaining to be filled.

Note that in general $Q(\bar{w}^k, x^*)$ converging to 0 does not necessarily guarantee that $\bar{w}^k$ converges to a solution x^*. Under additional condition, such as when $g(x)$ is strictly convex, the convergence to x^* can be shown.

Corollary 5.3.7. *If $g(x)$ is strictly convex, then $\lim_{k\to\infty} \mathbb{E}[\|\bar{w}^k - x^*\|] = 0$ in the setting of Corollary 5.3.5.*

Proof. Let $M(r) := \min_{\|x'-x^*\|=r} g(x') - g(x^*) - \langle \nabla g(x^*), x' - x^* \rangle$. Since $g(x)$ is strictly convex, $M(r) > 0$ as $r > 0$. Notice that for any $z' \in \mathcal{Z}$,

$$\begin{aligned} Q(x'; x^*) &= \langle H(x^*), x' - x^* \rangle + g(x') - g(x^*) \\ &= \langle H(x^*) + \nabla g(x^*), x - x^* \rangle + g(x') - g(x^*) - \langle \nabla g(x^*), x' - x^* \rangle \\ &\geq g(x') - g(x^*) - \langle \nabla g(x^*), x' - x^* \rangle. \end{aligned}$$

With Corollary 5.3.5, we get $\lim_{k\to\infty} \mathbb{E}[M(\|\bar{w}^k - x^*\|)] = 0$.

Note $\frac{M(r)}{r} = \min_{\|\theta\|=1} \frac{g(x^*+r\theta)-g(x^*)-\langle \nabla g(x^*), r\theta \rangle}{r}$ is increasing with respect to r. So, given any $\epsilon > 0$, $M(r) \geq r\frac{M(\epsilon)}{\epsilon}$ for any $r > \epsilon$, which implies

$$\mathbb{E}[M(\|\bar{w}^k - x^*\|)] \geq \mathbb{E}[M(\|\bar{w}^k - x^*\|)\mathbf{1}_{\|\bar{w}^k - x^*\| \geq \epsilon}] \geq \frac{M(\epsilon)}{\epsilon} \mathbb{E}[\|\bar{w}^k - x^*\| \mathbf{1}_{\|\bar{w}^k - x^*\| \geq \epsilon}].$$

Therefore, $\lim_{k\to\infty} \mathbb{E}[\|\bar{w}^k - x^*\| \mathbf{1}_{\|\bar{w}^k - x^*\| \geq \epsilon}] = 0$ and

$$\overline{\lim_{k\to\infty}} \mathbb{E}[\|\bar{w}^k - x^*\|] = \lim_{k\to\infty} \mathbb{E}[\|\bar{w}^k - x^*\| \mathbf{1}_{\|\bar{w}^k - x^*\| \geq \epsilon}] + \overline{\lim_{k\to\infty}} \mathbb{E}[\|\bar{w}^k - x^*\| \mathbf{1}_{\|\bar{w}^k - x^*\| < \epsilon}] \leq \epsilon.$$

Since $\epsilon > 0$ can be chosen arbitrarily, we have $\lim_{k\to\infty} \mathbb{E}[\|\bar{w}^k - x^*\|] = 0$. □

5.4 Finite-Sum Constrained Finite-Sum Optimization

In this section, we introduce an application for which the proposed SAVREP and SAVREP-m can be applied to. Consider the following problem:

$$\begin{array}{lll} (P) & \min & \sum_{i=1}^{m_2} g_i(x) \\ & \text{s.t.} & \sum_{j=1}^{m_1} h_j(x) \leq 0 \\ & & x \in \mathcal{X}. \end{array} \tag{5.4.35}$$

While it is not uncommon to formulate the objective function as finite-sum in machine learning research, the specific finite-sum structure of inequality constraints given in (5.4.35) is also found in applications such as empirical risk minimization and Neyman-Pearson classification [94]. Previous research [4, 52, 51] has developed level-set methods for solving (5.4.35). In particular, [51] proposed to reformulate the level-set subproblem into saddle-point problem and solve it with variance-reduced method [74].

5.4.1 A noise-free VI reformulation

In this paper, we propose to solve (5.4.35) through its Lagrangian dual formulation, which is equivalently a saddle point problem with a special structure that is suitable for applying the accelerated variance reduced method SAVREP-m. In our discussion, we assume $g_i(x)$ is convex for all $i = 1, ..., m_2$, $h_j(x) = (h_{j,1}(x), \cdots, h_{j,\ell}(x))^\top \in \mathbb{R}^\ell$ and $h_{j,s}(x)$ is convex in x for all $j = 1, ..., m_1$ and $s = 1, ..., \ell$, and $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set. The corresponding saddle point reformulation of (5.4.35) solves the following:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathbb{R}^\ell_+} \quad L(x;y) := \sum_{i=1}^{m_2} g_i(x) + \sum_{j=1}^{m_1} y^\top h_j(x), \tag{5.4.36}$$

where $L(x;y)$ defines the Lagrangian function of (P). The partial gradients of the Lagrangian function are given by:

$$\begin{cases} \nabla_x L(x;y) &= \sum_{i=1}^{m_2} \nabla g_i(x) + \sum_{j=1}^{m_1} (Jh_j(x))^\top y \\ \nabla_y L(x;y) &= \sum_{j=1}^{m_1} h_j(x). \end{cases}$$

Denote $\mathcal{Y} := \mathbb{R}^\ell_+$, then the optimality condition for (5.4.36) is the following VI problem:

Find $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\begin{pmatrix} \nabla_x L(x^*;y^*) \\ -\nabla_y L(x^*;y^*) \end{pmatrix}^\top \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \geq 0, \quad \text{for all } (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{5.4.37}$$

Or simply:

Find $z^* \in \mathcal{Z}$ such that

$$\langle F(z^*), z - z^* \rangle \geq 0, \text{ for all } z \in \mathcal{Z},$$

where we let $z := (x; y)$, $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, and

$$
\begin{aligned}
(5.4.38) \qquad F(z) \quad &:= \quad \sum_{j=1}^{m_1} \begin{pmatrix} (Jh_j(x))^\top y \\ -h_j(x) \end{pmatrix} + \sum_{i=1}^{m_2} \begin{pmatrix} \nabla g_i(x) \\ 0 \end{pmatrix} = \begin{pmatrix} \nabla_x L(x;y) \\ -\nabla_y L(x;y) \end{pmatrix} \\
&= \quad \sum_{j=1}^{m_1} H_j(z) + \sum_{i=1}^{m_2} \nabla g_i(z).
\end{aligned}
$$

Note that we have transformed the original finite-sum constrained finite-sum optimization problem (5.4.35) into solving a VI problem with the operator defined in (5.4.38), and such $F(z)$ indeed takes the form of (5.1.1), which consists of a finite-sum general VI mappings $\sum_{j=1}^{m_1} \begin{pmatrix} (Jh_j(x))^\top y \\ -h_j(x) \end{pmatrix}$ and a finite-sum gradient mappings $\sum_{i=1}^{m_2} \begin{pmatrix} \nabla g_i(x) \\ 0 \end{pmatrix}$. We caution that the *variables* x, y used in these expressions should not be confused with the *sequences* $\{x^k\}, \{y^k\}$ presented in the update (5.2.7) or (5.3.22). The former corresponds to the (dual) variables in the original optimization problem, while the latter is general VI variables. We use z as the variable in the VI reformulation (5.4.38) to distinguish between the two.

Note that the Jacobian of $\begin{pmatrix} \nabla g_i(x) \\ 0 \end{pmatrix}$ is $\begin{bmatrix} \nabla^2 g_i(x) & 0 \\ 0 & 0 \end{bmatrix}$, which is positive semidefinite since each $f_i(x)$ is assumed to be convex. On the other hand, the Jacobian of $\begin{pmatrix} (Jh_j(x))^\top y \\ -h_j(x) \end{pmatrix}$ is

$$
\begin{pmatrix} \sum_{s=1}^{\ell} y_s \nabla^2 h_{j,s}(x) & (Jh_j(x))^\top \\ -Jh_j(x) & 0_{\ell \times \ell} \end{pmatrix}.
$$

Since $y_s \geq 0$ and $h_{j,s}(x)$ is convex, the above Jacobian matrix is also positive semidefinite. Therefore, we can conclude that the operator $F(z)$ in the VI reformulation is indeed monotone.

While the efficiency of the variance reduced algorithms for optimization is now commonly recognized when the total number m_2 of functions $g_i(x)$ in the summation is large, it is also reasonable to apply similar variance reduced techniques for estimating the constraint functions $h_j(x)$ when the total number m_1 in the summation is large, as it can be costly to evaluate all these constraint functions (or their Jacobians) in each iteration. Problem (P) in (5.4.35) describes exactly such a situation, and by reformulating the original problem into a finite-sum VI with the special structure (5.4.38), the proposed SAVREP-m in Section 5.3 can be applied. It incorporates variance reduction into the update process for both finite-sum gradient mappings and finite-sum VI mappings, where the latter is attributed to the (Jacobians) of the constraints $h_i(x)$ and the corresponding dual variable y. Note that since the dual variable y is constrained to be non-negative, Assumption 5.3.2 in general does not

hold in our VI reformulation with the operator (5.4.38), where the constraint is given by $\mathcal{Z} := \mathcal{X} \times \mathbb{R}^{\ell}_{+}$. However, it is merely a convenient assumption for deriving the gradient complexity guarantee in our analysis, while in practice it makes sense to set a large enough diameter constant that contains the optimal dual variable y^* and perform projections onto the bounded constrained set instead. As we will show in Section 5.5, the improved gradient complexities due to applying variance reduction respectively to the general VI mapping and gradient mapping are indeed observed regardless of the boundedness of the constraint sets in our most general VI reformulation (5.4.37).

Alternatively, one can also apply SAVREP proposed in Section 5.2, which solves a strongly monotone VI instead. While the operator (5.4.38) in our VI reformulation is merely monotone, it can be easily transformed to a strongly monotone VI by considering the following *approximated* VI problem with the perturbed operator:

$$F_{\mu}(z) := F(z) + \mu z, \tag{5.4.39}$$

which is strongly monotone with $\mu > 0$ with $F(z)$ defined in (5.4.38). Note that SAVREP only requires $H(z) = \sum_{j=1}^{m_1} H_j(z)$ to be strongly monotone, so the perturbation term μz can be associated to $H_j(z)$ for arbitrary $j = 1, 2, .., m_1$. In particular, we can construct the variance reduced gradient estimators in (5.2.12) as:

$$\hat{H}(z^{k+0.5}) := H(w^k) + H_{\xi_k}(z^{k+0.5}) - H_{\xi_k}(w^k) + \mu z^{k+0.5},$$

where ξ_k randomly samples from $j = 1, 2, .., m_1$ and $H_j(\cdot)$ is defined in (5.4.38). The counterpart for $\tilde{\nabla} g(z^k)$ remains unchanged from (5.2.13).

To ensure that the solution obtained from the VI associated with $F_{\mu}(z)$ serves as a good approximated solution to the original VI when μ is small, let us also introduce the following *error bound* assumption:

Assumption 5.4.1 (Error Bound)**.** *Let $F(z)$ be monotone and $\mu > 0$. Denote $z^*(\mu)$ as the solution to the VI problem with operator $F_{\mu}(z)$, namely:*

$$\langle F_{\mu}(z^*(\mu)), z - z^*(\mu) \rangle \geq 0, \quad \forall z \in \mathcal{Z},$$

and let z^* be a solution to the VI with operator $F(z)$. There exist constants $\theta \in (0, 1]$, c_1, $c_2 > 0$ such that for all $0 < \mu < c_1$, the following holds:

$$\|z^*(\mu) - z^*\| \leq c_2 \cdot \mu^{\theta}.$$

Assumption 5.4.1 ensures that by solving an approximated solution to the strongly monotone VI with operator $F_{\mu}(z)$, we are able to use the exact same solution as the approximated

solution to the original monotone VI with $F(z)$, which in turn solves (5.4.35). A similar error bound assumption for convex-concave saddle point problem is also discussed in Assumption 5.1 in [25], where they further showed that the assumption holds with $\theta = 1$ for quadratic saddle point functions with bilinear coupling. In theory, taking a perturbation parameter $\mu = O(\epsilon^{\frac{1}{\theta}})$ while applying SAVREP to obtain a $\frac{\epsilon}{2}$-solution z^k will guarantee the same z^k to be an ϵ-solution to the operator $F(z)$. In practice, the single-loop structure of SAVREP makes it easier to implement compared to its monotone variant SAVREP-m.

5.4.2 A stochastic zeroth-order approach

In this section, we consider the same problem (P) in (5.4.35) but assume that only the unbiased noisy estimations of each of the function values $g_i(x)$ and constraint function values $h_{j,s}(x)$ are accessible, denoted by $g_i'(x)$ and $h_{j,s}'(x)$ respectively. Under this assumption, the gradients of the functions $\nabla g_i(x)$,$\nabla h_{j,s}(x)$ are not directly available, and the methods applied to problems of this type are often referred to as *derivation-free* or *zeroth-order* methods. There have been developments in the context of optimization [72, 87, 17, 88, 46], as well as in the context of saddle point problems [99, 101, 54, 85, 56, 26].

In the following discussion, we present a zeroth-order approach based on the saddle point reformulation (5.4.36). By applying the randomized smoothing approach [72], we can replace the gradients $\nabla g_i(x)$, $\nabla h_{j,s}(x)$ required in the VI operator (5.4.38) with the *stochastic zeroth-order gradients*. The resulting VI operator then serves as the stochastic estimators used in the update of SAVREP (5.2.7), and we shall derive the stochastic bounds in (5.2.8)-(5.2.9) in terms of the parameters involved in the construction of these stochastic zeroth-order gradients.

Let us first state the assumptions for the stochastic oracles $g_i'(x)$, $h_{j,s}'(x)$:

(5.4.40)

$$\begin{cases} \mathbb{E}'\left[g_i'(x)\right] = g_i(x), \\ \mathbb{E}'\left[\nabla g_i'(x)\right] = \nabla g_i(x), \\ \mathbb{E}'\left[\left\|\nabla g_i'(x) - \nabla g_i(x)\right\|^2\right] \le \varsigma_g^2, \end{cases} \qquad \begin{cases} \mathbb{E}'[h_{j,s}'(x)] = h_{j,s}(x), \\ \mathbb{E}'\left[|h_{j,s}'(x) - h_{j,s}(x)|^2\right] \le \varpi^2, \\ \mathbb{E}'[\nabla h_{j,s}'(x)] = \nabla h_{j,s}(x), \\ \mathbb{E}'\left[\left\|\nabla h_{j,s}'(x) - \nabla h_{j,s}(x)\right\|^2\right] \le \varsigma_h^2. \end{cases}$$

Note that we have used $\mathbb{E}'[\cdot]$ as the expectation taken for the stochastic oracle, suppressing the notation of random variable for simplicity. In addition, we assume that $g_i(x)$, $\nabla g_i(x)$, $h_{j,s}(x)$, $\nabla h_{j,s}(x)$ are Lipschitz continuous with constants $M_{i,g}, L_{i,g}, M_{j,s,h}, L_{j,s,h}$ respectively.

Given a function $g(x)$, the corresponding smoothing function with parameter φ can be

obtained by taking the expectation of random samples taken from the uniform distribution U_b on a Euclidean ball B in $\mathbb{R}^n$, defined as following:

$$g_\varphi(x) := \mathbb{E}_{u\sim U_b}\left[g(x+\varphi u)\right] = \frac{1}{\alpha(n)}\int_B g(x+\varphi u)du,$$

where $\alpha(n)$ is the volume of the unit ball B. The above smoothing function $g_\varphi(x)$ is then continuously differentiable regardless of the continuity of the original function $g(x)$. We summarize some properties of the smoothing function and its gradient in the next lemma, which can also be found in the literature of zeroth-order methods.

Lemma 5.4.2. *The smoothing function $g_\varphi(x)$ is continuously differentiable. Denote U_{S_p} as the uniform distribution on the unit sphere S_p in $\mathbb{R}^n$. The gradient $\nabla g_\varphi(x)$ can be expressed as the following:*

$$\nabla g_\varphi(x) := \mathbb{E}_{u\sim U_{S_p}}\left[\frac{n}{\varphi}g(x+\varphi u)u\right] = \mathbb{E}_{u\sim U_{S_p}}\left[\frac{n}{\varphi}\left(g(x+\varphi u)-g(x)\right)u\right].$$

Furthermore, if $g(x)$ is also differentiable, then the following bounds hold:

$$\|\nabla g_\varphi(x)-\nabla g(x)\| \le \frac{\varphi n L}{2}, \tag{5.4.41}$$

$$\mathbb{E}_{u\sim U_{S_p}}\left[\left\|\frac{n}{\varphi}\left(g(x+\varphi u)-g(x)\right)u\right\|^2\right] \le 2n\|\nabla g(x)\|^2 + \frac{\varphi^2 n^2 L^2}{2}, \tag{5.4.42}$$

where L is the Lipschitz constant of $\nabla g(x)$.

The proof of Lemma 5.4.2 can be found in the literature, and we refer the interested readers to [87] (Lemma 4.4) and [20] (Propositions 2.7.5 and 2.7.6) for the details.

Based on the properties in Lemma 5.4.2, we now define the *stochastic zeroth-order gradient* as the following:

$$\begin{aligned} G'_{i,\varphi}(x,u) &:= \frac{n}{\varphi}\left(g'_i(x+\varphi u)-g'_i(x)\right)u, \\ H'_{j,s,\varphi}(x,u) &:= \frac{n}{\varphi}\left(h'_{j,s}(x+\varphi u)-h'_{j,s}(x)\right)u, \end{aligned}$$

where $u \sim U_{S_p}$, and we have replaced the function values $g(x)$ with the corresponding stochastic oracles $g'_i(x)$ and $h'_{j,s}(x)$ for each of the function. Note that when evaluating the stochastic zeroth-order gradient $G'_{i,\varphi}(x,u)$ ($H'_{j,s,\varphi}(x,u)$), we use the *same* random variable ξ_g (ξ_h) to evaluate the stochastic function estimator $g'_i(\cdot)$ ($h'_{j,s}(\cdot)$) at $x+\varphi u$ and x respectively. The dependency on the random variables is suppressed for simplicity. The next corollary states that the stochastic zeroth-order gradient is an unbiased estimator of the gradient of the smoothing function $\nabla g_\varphi(x)$ with bounded variance.

Corollary 5.4.3. *The stochastic zeroth-order gradients are unbiased with respect to the gradient of the smoothing function with bounded variance:*

$$\mathbb{E}'_u\left[G'_{i,\varphi}(x,u)\right] = \nabla g_{i,\varphi}(x), \quad \mathbb{E}'_u\left[H'_{j,s,\varphi}(x,u)\right] = \nabla h_{j,s,\varphi}(x),$$

and

$$\mathbb{E}'_u\left[\left\|G'_{i,\varphi}(x,u) - \nabla g_{i,\varphi}(x)\right\|^2\right] \le \tilde{\varsigma}_g^2 := 2n\left(M_{i,g}^2 + \varsigma_g^2\right) + \frac{\varphi^2 n^2 L_{i,g}^2}{2}, \tag{5.4.43}$$

$$\mathbb{E}'_u\left[\left\|H'_{j,s,\varphi}(x,u) - \nabla h_{j,s,\varphi}(x)\right\|^2\right] \le \tilde{\varsigma}_h^2 := 2n\left(M_{j,s,h}^2 + \varsigma_h^2\right) + \frac{\varphi^2 n^2 L_{j,s,h}^2}{2}. \tag{5.4.44}$$

Proof. *See Appendix 5.7.6.* □

Now we can replace the gradient mappings in the operator $F(z)$ of the VI reformulation (5.4.38) with the stochastic zeroth-order gradients:

$$F'(z) := \sum_{j=1}^{m_1}\begin{pmatrix} H'_{j,\varphi}(x,u)y \\ -h'_j(x)\end{pmatrix} + \sum_{i=1}^{m_2}\begin{pmatrix} G'_{i,\varphi}(x,u) \\ 0\end{pmatrix} := \sum_{j=1}^{m_1} H'_j(z) + \sum_{i=1}^{m_2}\nabla g'_i(z), \tag{5.4.45}$$

where $H'_{j,\varphi}(x,u) := (H'_{j,1,\varphi}(x,u), H'_{j,2,\varphi}(x,u), ..., H'_{j,\ell,\varphi}(x,u))$ is a matrix with column vectors being the stochastic zeroth-order gradient $H'_{j,s,\varphi}(x,u)$ for $s = 1,2,...,\ell$, and $h'_j(x) := (h'_{j,1}(x), ..., h'_{j,\ell}(x))^\top$. By constructing the stochastic zeroth-order operator $F'(z)$ as in (5.4.45), the proposed SAVREP (5.2.7) is readily applicable to the VI reformulation of problem (P) when the function value estimations are noisy. We conclude this section by summarizing the corresponding stochastic bounds in the forms of (5.2.8)-(5.2.9).

Corollary 5.4.4. *Let $H_j(z), \nabla g_i(z)$ be defined in (5.4.38) and $H'_j(z), \nabla g'_i(z)$ be defined in (5.4.45). Furthermore, denote K as the total number of iterations performed by SAVREP and define $D_y := \max_{0\le k\le K}\|y\|$, then we have the following stochastic bounds hold for all $z \in \{z^k\}_{0\le k\le K}$:*

$$\begin{aligned}
&\left\|H_j(z) - \mathbb{E}'[H'_j(z)]\right\| \le \delta_h := \frac{\varphi n D_y}{2}\sqrt{\sum_{s=1}^{\ell} L_{j,s,h}^2}, \\
&\mathbb{E}'\left[\left\|H'_j(z) - \mathbb{E}'\left[H'_j(z)\right]\right\|^2\right] \le \sigma_h^2 := \ell\left(\tilde{\varsigma}_h^2 D_y^2 + \varpi^2\right), \\
&\left\|\nabla g_i(z) - \mathbb{E}'\left[\nabla g'_i(z)\right]\right\| \le \delta_g := \frac{\varphi n L_{i,g}}{2}, \\
&\mathbb{E}'\left[\left\|\nabla g'_i(z) - \mathbb{E}'\left[\nabla g'_i(z)\right]\right\|^2\right] \le \sigma_g^2 := \tilde{\varsigma}_g^2.
\end{aligned}$$

Proof. *See Appendix 5.7.7.* □

5.5 Numerical Experiments

In this section, we evaluate the numerical performance of SAVREP and SAVREP-m by using the same example as in [51], which is a Neyman-Pearson classification problem [94] formulated as

$$\min_{\|\mathbf{x}\|_2 \le \lambda} \frac{1}{n_0} \sum_{j=1}^{n_0} \phi\left(\mathbf{x}^\top \xi_{0j}\right), \text{ s.t. } \frac{1}{n_1} \sum_{j=1}^{n_1} \phi\left(-\mathbf{x}^\top \xi_{1j}\right) \le r_1, \tag{5.5.46}$$

where ϕ is the loss function, defined as smoothed hinge loss function in the experiment for SAVREP and logistic loss function in the experiment for SAVREP-m. The dataset is the rcv1 training data set from LIBSVM library with $20,242$ data points with $n_0 = 10,491$ and $n_1 = 9,751$ and a dimension of $47,236$.

5.5.1 SAVREP

In this experiment, the loss function is defined as

$$\phi(t) = \begin{cases} \frac{1}{2} - t, & t \le 0, \\ \frac{1}{2}(1-t)^2, & 0 < t \le 1, \\ 0, & t > 1, \end{cases}$$

and we focus on the perturbed problem 5.4.39. The parameters are set as $\lambda = 5$ and $r_1 = 0.1$, and the perturbation is set as $\mu = 10^{-5}, 10^{-10}$ respectively. We compare the performance of SAVREP with extragradient with variance reduction (EVR) [2]. Both of the methods use the mini-batch with a batch size of 100 to get the stochastic gradient estimators. We tune τ for EVR methods and α and γ for SAVREP. To give a fair comparison, for all the parameters we tune, we select learning rates from the set $\left\{10^{-k}, 2\times 10^{-k}, 4\times 10^{-k}, 8\times 10^{-k} : k \in \mathbb{Z}\right\}$ times the parameter settings for theoretical analysis in their corresponding paper. We use both the distance from the iterates to the optimal solution (solved by CVX mosek) and the norm of $H(x) + \nabla g(x)$ as the performance measure. The results are shown in Figure 5.1 ($\mu = 10^{-5}$) and Figure 5.2 ($\mu = 10^{-10}$), with left plots showing distance to the optimal solution and right plots showing the norm of $H(x)+\nabla g(x)$. In these experiments, SAVREP shows faster convergence than EVR. Furthermore, the distance to the optimal solution is shorter for smaller perturbation $\mu = 10^{-10}$, which is in line with expectation.

5.5.2 SAVREP-m

In this experiment, we test SAVREP-m on the same problem (5.5.46), using the VI formulation without perturbation (5.4.38). The parameter tuning is similar to the previ-

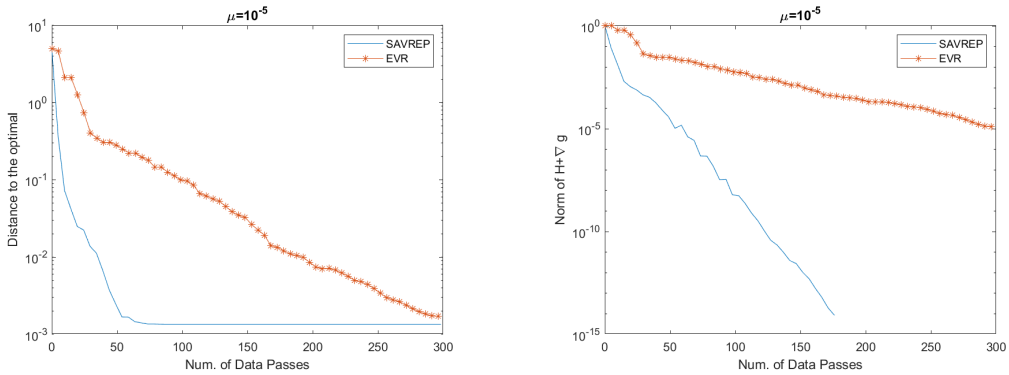

Figure 5.1: Convergence of SAVREP under perturbation $\mu = 10^{-5}$

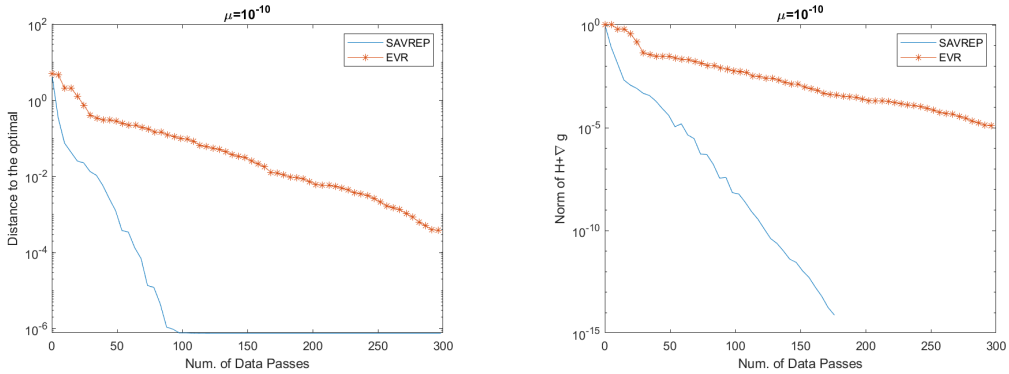

Figure 5.2: Convergence of SAVREP under perturbation $\mu = 10^{-10}$

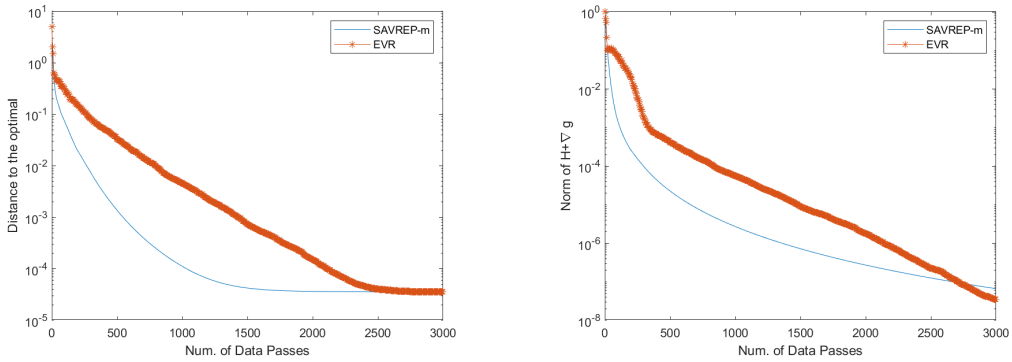

Figure 5.3: Convergence of SAVREP-m: distance to optimal solution (left) and norm of $H + \nabla g$ (right)

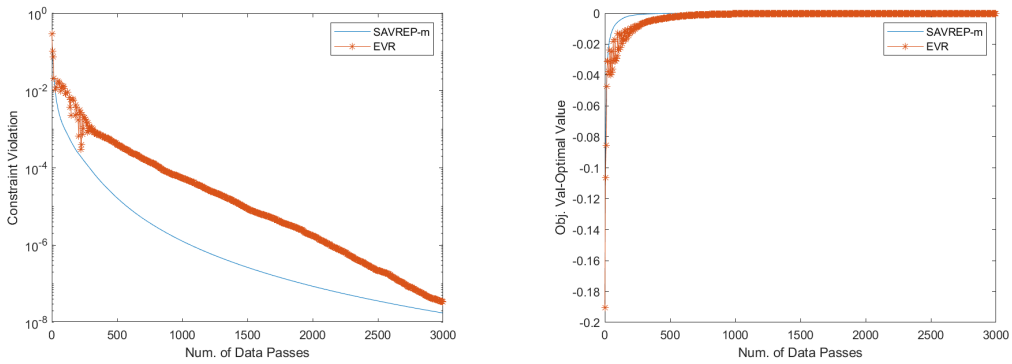

Figure 5.4: Convergence of SAVREP-m: constraint violation (left) and objective function gap (right)

ous experiment, while the loss function is defined as the logistic loss function, i.e. $\phi(t) = \log(1+\exp(-t))$, with $\lambda = 5$ and $r_1 = 0.4$. The convergence in terms of distance to optimal solution and norm of $H(x) + \nabla g(x)$ are shown in Figure 5.3. In addition, the constraint violation and objective function gap are shown in Figure 5.4. These results demonstrate the sublinear convergence rate of SAVREP-m, which is consistent with the theoretical guarantee derived in Section 5.3.1. On the other hand, EVR in this experiment shows a linear convergence rate similar to that in the previous experiment with perturbation. Note that in the second experiment the problem may still be (locally) strongly monotone after reformulation even without perturbation. While EVR does not require the strongly monotone modulus in its algorithm and reflects linear convergence automatically, we note that SAVREP-m *explicitly* assumes the problem to be merely monotone by using diminishing step sizes, as shown in Corollary 5.3.5. The specific design is necessary for SAVREP-m to guarantee a faster sublinear rate given the composite VI structure (5.2.6), at the cost of not being able to converge linearly when the problem is actually strongly monotone. We remark that the purpose of this experiment is to demonstrate the convergence behavior of SAVREP-m, and in practice it makes sense to apply SAVREP instead with estimated strong monotonicity modulus.

5.6 Conclusion

In this chapter, we propose two stochastic variance reduced schemes, SAVREP and SAVREP-m, for solving an extended class of finite-sum VI with strongly monotone operator and monotone operator respectively. The operator consists of the sum of a general VI mapping and a gradient mapping, both with finite-sum structure. By exploiting this special

structure and applying variance reduction techniques developed in the literature, we show that both schemes admit improved gradient complexities, compared to existing variance reduction algorithms proposed for general finite-sum VI. In addition, we consider a more general stochastic setup in both proposed schemes, where the stochastic oracles of noisy mappings are adopted in the updates, and derive the corresponding stochastic bounds in the results of our complexity analysis. We show that an application of finite-sum optimization with finite-sum inequality constraints can be reformulated into the finite-sum VI of the special structure discussed in this paper, where the proposed schemes can be readily applied to. Finally, we note that while the established gradient complexity results match the optimal complexities in terms of problem constants in optimization (i.e. the constants related to the gradient mapping), the gap between the current lower bound established for finite-sum VI remains unfilled. It is still an open question whether the upper bound or the lower bound can be improved to match the other, and we leave it to future works.

5.7 Appendix: Proofs of technical results

5.7.1 Proof of Lemma 5.2.1

The optimality condition at $x^{k+0.5}$ yields

$$\langle \gamma \left(H'(w^k) + \tilde{\nabla} g'(y^k) \right) + x^{k+0.5} - \bar{x}^k, x - x^{k+0.5} \rangle \geq 0, \quad \forall x \in \mathcal{Z}, \tag{5.7.47}$$

and the optimality condition at x^{k+1} yields

$$\langle \gamma \left(\hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k) \right) + x^{k+1} - \bar{x}^k, x - x^{k+1} \rangle \geq 0, \quad \forall x \in \mathcal{Z}. \tag{5.7.48}$$

From (5.7.48), use the expression of $\bar{x}^k$:

$$\begin{aligned} & \frac{1}{2} \left(\|x^{k+1} - x\|^2 + (1-p_1)\|x^{k+1} - x^k\|^2 - (1-p_1)\|x^k - x\|^2 \right. \\ & \left. + p_1 \|x^{k+1} - w^k\|^2 - p_1 \|w^k - x\|^2 \right) \\ = \; & (1-p_1)\langle x^{k+1} - x^k, x^{k+1} - x \rangle + p_1 \langle x^{k+1} - w^k, x^{k+1} - x \rangle \\ \leq \; & \gamma \langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+1} \rangle, \end{aligned} \tag{5.7.49}$$

where

$$\begin{aligned} & \gamma \langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+1} \rangle \\ = \; & \gamma \langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+0.5} \rangle + \gamma \langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x^{k+0.5} - x^{k+1} \rangle \\ = \; & \gamma \langle H(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+0.5} \rangle + \gamma \langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5} \rangle \\ & + \gamma \langle H'(w^k) + \tilde{\nabla} g'(y^k), x^{k+0.5} - x^{k+1} \rangle + \gamma \langle \hat{H}'(x^{k+0.5}) - H'(w^k), x^{k+0.5} - x^{k+1} \rangle. \end{aligned} \tag{5.7.50}$$

The third term in the above inequality can be bounded by using (5.7.47) with $x = x^{k+1}$:

$$\begin{aligned}
& \gamma\langle H'(w^k) + \tilde{\nabla} g'(y^k), x^{k+0.5} - x^{k+1}\rangle \leq \langle x^{k+0.5} - \bar{x}^k, x^{k+1} - x^{k+0.5}\rangle \\
= \; & (1-p_1)\langle x^{k+0.5} - x^k, x^{k+1} - x^{k+0.5}\rangle + p_1\langle x^{k+0.5} - w^k, x^{k+1} - x^{k+0.5}\rangle \\
= \; & \frac{1}{2}\Big(-\|x^{k+1} - x^{k+0.5}\|^2 + (1-p_1)\|x^{k+1} - x^k\|^2 - (1-p_1)\|x^{k+0.5} - x^k\|^2 \\
& + p_1\|x^{k+1} - w^k\|^2 - p_1\|x^{k+0.5} - w^k\|^2\Big),
\end{aligned}$$

while the fourth term can be bounded by:

$$\begin{aligned}
& \gamma\langle \hat{H}'(x^{k+0.5}) - H'(w^k), x^{k+0.5} - x^{k+1}\rangle = \gamma\langle H'_{\xi_k}(x^{k+0.5}) - H'_{\xi_k}(w^k), x^{k+0.5} - x^{k+1}\rangle \\
\leq \; & \frac{\gamma^2}{2}\|H'_{\xi_k}(x^{k+0.5}) - H'_{\xi_k}(w^k)\|^2 + \frac{1}{2}\|x^{k+0.5} - x^{k+1}\|^2.
\end{aligned}$$

Combine the above two inequalities with (5.7.50) and use $\mathbb{E}_{k_1}[\cdot] := \mathbb{E}_{\xi_k}[\cdot|x^k, w^k]$:

$$\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma\langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+1}\rangle\right] \\
\leq \; & \mathbb{E}_{k_1}\left[\gamma\langle H(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+0.5}\rangle\right] + \mathbb{E}_{k_1}\left[\gamma\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right] \\
& + \frac{1}{2}\mathbb{E}_{k_1}\left[\gamma^2\|H'_{\xi_k}(x^{k+0.5}) - H'_{\xi_k}(w^k)\|^2 + (1-p_1)\|x^{k+1} - x^k\|^2\right. \\
& \left. -(1-p_1)\|x^{k+0.5} - x^k\|^2 + p_1\|x^{k+1} - w^k\|^2 - p_1\|x^{k+0.5} - w^k\|^2\right].
\end{aligned}$$

(5.7.51)

Let us bound the term $\mathbb{E}_{k_1}\left[\gamma\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right]$. Note that with $\mathbb{E}_{k_1}[\cdot]$, $x^{k+0.5}$ is deterministic and $\mathbb{E}_{k_1}\left[\hat{H}'(x^{k+0.5})\right] = H'(x^{k+0.5})$. Therefore,

$$\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right] = \gamma\langle H'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle \\
\leq \; & \frac{\gamma}{2\mu_h}\left\|H'(x^{k+0.5}) - H(x^{k+0.5})\right\|^2 + \frac{\gamma\mu_h}{2}\|x - x^{k+0.5}\|^2.
\end{aligned}$$

(5.7.52)

Continue with (5.7.51):

$$\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma\langle \hat{H}'(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+1}\rangle\right] \\
\leq \; & \mathbb{E}_{k_1}\left[\gamma\langle H(x^{k+0.5}) + \tilde{\nabla} g'(y^k), x - x^{k+0.5}\rangle\right] + \frac{1}{2}\mathbb{E}_{k_1}\left[\gamma^2\|H'_{\xi_k}(x^{k+0.5}) - H'_{\xi_k}(w^k)\|^2\right] \\
& + \frac{1}{2}\mathbb{E}_{k_1}\left[(1-p_1)\|x^{k+1} - x^k\|^2 - (1-p_1)\|x^{k+0.5} - x^k\|^2 + p_1\|x^{k+1} - w^k\|^2\right. \\
& \left. -p_1\|x^{k+0.5} - w^k\|^2\right] + \frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h} + \frac{1}{2}\mathbb{E}_{k_1}\left[\gamma\mu_h\|x^{k+0.5} - x\|^2\right].
\end{aligned}$$

Combine with (5.7.49):

$$
\begin{aligned}
& \frac{1}{2}\mathbb{E}_{k_1}\left[\|x^{k+1}-x\|^2-(1-p_1)\|x^k-x\|^2-p_1\|w^k-x\|^2+(1-p_1)\|x^{k+0.5}-x^k\|^2\right] \\
& +\mathbb{E}_{k_1}\left[\gamma\langle H(x^{k+0.5})+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[\gamma^2\|H'_{\xi_k}(x^{k+0.5})-H'_{\xi_k}(w^k)\|^2-p_1\|x^{k+0.5}-w^k\|^2\right]+\frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h} \\
& +\frac{1}{2}\mathbb{E}_{k_1}\left[\gamma\mu_h\|x^{k+0.5}-x\|^2\right] \\
\leq\ & \frac{1}{2}(2\gamma^2L_h^2-p_1)\|x^{k+0.5}-w^k\|^2+\frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h} \\
& +\frac{1}{2}\mathbb{E}_{k_1}\left[\gamma\mu_h\|x^{k+0.5}-x\|^2\right]+\gamma^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right].
\end{aligned}
$$

The last inequality is due to the following relation:

$$
\begin{aligned}
& \mathbb{E}_{k_1}\left[\|H'_{\xi_k}(x^{k+0.5})-H'_{\xi_k}(w^k)\|^2\right] \\
\leq\ & \mathbb{E}_{k_1}\left[2\|H_{\xi_k}(x^{k+0.5})-H_{\xi_k}(w^k)\|^2\right] \\
& +\mathbb{E}_{k_1}\left[2\left(\|H'_{\xi_k}(x^{k+0.5})-H_{\xi_k}(x^{k+0.5})\|+\|H'_{\xi_k}(w^k)-H_{\xi_k}(w^k)\|\right)^2\right] \\
\leq\ & 2L_h^2\|x^{k+0.5}-w^k\|^2+2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right].
\end{aligned}
$$

(5.7.53)

Rearrange the terms with the strong monotonicity of $H(\cdot)$:

$$
\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma\langle H(x)+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle+\gamma\mu_h\|x^{k+0.5}-x\|^2+\frac{1}{2}(p_1-2\gamma^2L_h^2)\|x^{k+0.5}-w^k\|^2\right] \\
\leq\ & \mathbb{E}_{k_1}\left[\gamma\langle H(x^{k+0.5})+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle+\frac{1}{2}(p_1-2\gamma^2L_h^2)\|x^{k+0.5}-w^k\|^2\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[-\|x^{k+1}-x\|^2+(1-p_1)\|x^k-x\|^2+p_1\|w^k-x\|^2-(1-p_1)\|x^{k+0.5}-x^k\|^2\right] \\
& +\frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h}+\frac{1}{2}\mathbb{E}_{k_1}\left[\gamma\mu_h\|x^{k+0.5}-x\|^2\right]+\gamma^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right],
\end{aligned}
$$

which gives us

$$
\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma\langle H(x)+\tilde{\nabla}g'(y^k), x^{k+0.5}-x\rangle\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[(1-p_1)\|x^k-x\|^2+p_1\|w^k-x\|^2-\|x^{k+1}-x\|^2\right] \\
& -\frac{1}{2}\mathbb{E}_{k_1}\left[(p_1-2\gamma^2L_h^2)\|x^{k+0.5}-w^k\|^2\right]-\frac{1}{2}\mathbb{E}_{k_1}\left[\gamma\mu_h\|x^{k+0.5}-x\|^2\right] \\
& -\frac{1}{2}\left(1-p_1\right)\|x^{k+0.5}-x^k\|^2+\frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h}+\gamma^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[(1-p_1)\|x^k-x\|^2+p_1\|w^k-x\|^2-\|x^{k+1}-x\|^2\right] \\
& -\frac{1}{2}\mathbb{E}_{k_1}\left[(p_1-2\gamma^2L_h^2)\|x^{k+0.5}-w^k\|^2\right]-\mathbb{E}_{k_1}\left[\frac{1}{4}\gamma\mu_h\|x^k-x\|^2-\frac{1}{2}\gamma\mu_h\|x^{k+0.5}-x^k\|^2\right] \\
& -\frac{1}{2}\left(1-p_1\right)\mathbb{E}_{k_1}\left[\|x^{k+0.5}-x^k\|^2\right]+\frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h}+\gamma^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[(1-p_1-\frac{1}{2}\gamma\mu_h)\|x^k-x\|^2+p_1\|w^k-x\|^2-\|x^{k+1}-x\|^2\right] \\
& -\frac{1}{2}\left(p_1-2\gamma^2L_h^2\right)\mathbb{E}_{k_1}\left[\|x^{k+0.5}-w^k\|^2\right]-\frac{1}{2}\left(1-p_1-\gamma\mu_h\right)\mathbb{E}_{k_1}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& +\frac{\gamma\bar{\varepsilon}^2_{x^{k+0.5}}}{2\mu_h}+\gamma^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right],
\end{aligned}
\tag{5.7.54}
$$

completing the proof.

5.7.2 Proof of Lemma 5.2.2

Using the Lipchistz continuity of $g(\cdot)$:

$$
\begin{aligned}
g(v^{k+1}) &\leq g(y^k)+\langle\nabla g(y^k), v^{k+1}-y^k\rangle+\frac{L_g}{2}\|v^{k+1}-y^k\|^2 \\
&= g(y^k)+\langle\nabla g(y^k),(1-\alpha-\beta)v^k+\alpha x^{k+0.5}+\beta\bar{w}^k-y^k\rangle+\frac{L_g\alpha^2}{2}\|x^{k+0.5}-x^k\|^2 \\
&= (1-\alpha-\beta)\left(g(y^k)+\langle\nabla g(y^k), v^k-y^k\rangle\right)+\alpha\left(g(y^k)+\langle\nabla g(y^k), x^{k+0.5}-y^k\rangle\right) \\
&\quad +\beta\left(g(y^k)+\langle\nabla g(y^k),\bar{w}^k-y^k\rangle\right)+\frac{L_g\alpha^2}{2}\|x^{k+0.5}-x^k\|^2 \\
&\leq (1-\alpha-\beta)g(v^k)+\alpha\left(g(x)+\langle\nabla g(y^k), x^{k+0.5}-x\rangle\right) \\
&\quad +\beta\left(g(y^k)+\langle\nabla g(y^k),\bar{w}^k-y^k\rangle\right)+\frac{L_g\alpha^2}{2}\|x^{k+0.5}-x^k\|^2 \\
&= (1-\alpha-\beta)g(v^k)+\beta\left(g(y^k)+\langle\nabla g(y^k),\bar{w}^k-y^k\rangle\right)+\frac{L_g\alpha^2}{2}\|x^{k+0.5}-x^k\|^2 \\
&\quad +\alpha\left(g(x)+\langle\tilde{\nabla}g'(y^k), x^{k+0.5}-x\rangle+\langle\nabla g(y^k)-\tilde{\nabla}g(y^k), x^{k+0.5}-x\rangle\right) \\
&\quad +\alpha\langle\tilde{\nabla}g(y^k)-\tilde{\nabla}g'(y^k), x^{k+0.5}-x\rangle.
\end{aligned}
\tag{5.7.55}
$$

Let us first bound the last term $\alpha\langle\tilde{\nabla}g(y^k) - \tilde{\nabla}g'(y^k), x^{k+0.5} - x\rangle$ in the above inequality (5.7.55). By taking the expectation $\mathbb{E}_{k_2}[\cdot]$:

$$\begin{aligned}
& \alpha\mathbb{E}_{k_2}\left[\langle\tilde{\nabla}g(y^k) - \tilde{\nabla}g'(y^k), x^{k+0.5} - x\rangle\right] \\
\leq\ & \mathbb{E}_{k_2}\left[\frac{4\alpha}{\mu_h}\left\|\tilde{\nabla}g(y^k) - \tilde{\nabla}g'(y^k)\right\|^2\right] + \mathbb{E}_{k_2}\left[\frac{\alpha\mu_h}{16}\|x^{k+0.5} - x\|^2\right] \\
\leq\ & \mathbb{E}_{k_2}\left[\frac{4\alpha}{\mu_h}\left\|\tilde{\nabla}g(y^k) - \tilde{\nabla}g'(y^k)\right\|^2\right] + \mathbb{E}_{k_2}\left[\frac{\alpha\mu_h}{8}\|x^{k+0.5} - x^k\|^2\right] + \mathbb{E}_{k_2}\left[\frac{\alpha\mu_h}{8}\|x^k - x\|^2\right].
\end{aligned}$$

Note that

$$\begin{aligned}
& \mathbb{E}_{k_2}\left[\left\|\tilde{\nabla}g(y^k) - \tilde{\nabla}g'(y^k)\right\|^2\right] \\
\leq\ & \mathbb{E}_{k_2}\left[\left(\|\nabla g(\bar{w}^k) - \nabla g'(\bar{w}^k)\| + \|\nabla g_{\zeta_k}(\bar{w}^k) - \nabla g'_{\zeta_k}(\bar{w}^k)\| + \|\nabla g_{\zeta_k}(y^k) - \nabla g'_{\zeta_k}(y^k)\|\right)^2\right] \\
\leq\ & 2\bar{\rho}^2_{\bar{w}^k} + 2\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k} + \rho_{y^k})^2\right].
\end{aligned}$$

Also note that:

$$\begin{aligned}
& \mathbb{E}_{k_2}\left[\langle\nabla g(y^k) - \tilde{\nabla}g(y^k), x^{k+0.5} - x\rangle\right] = \mathbb{E}_{k_2}\left[\langle\nabla g(y^k) - \tilde{\nabla}g(y^k), x^{k+0.5} - x^k\rangle\right] \\
\leq\ & \mathbb{E}_{k_2}\left[\frac{\beta}{2\alpha L_g}\|\nabla g(y^k) - \tilde{\nabla}g(y^k)\|^2\right] + \mathbb{E}_{k_2}\left[\frac{\alpha L_g}{2\beta}\|x^{k+0.5} - x^k\|^2\right] \\
\leq\ & \frac{\beta}{\alpha}\left(g(\bar{w}^k) - g(y^k) - \langle\nabla g(y^k), \bar{w}^k - y^k\rangle\right) + \mathbb{E}_{k_2}\left[\frac{\alpha L_g}{2\beta}\|x^{k+0.5} - x^k\|^2\right]. \qquad (5.7.56)
\end{aligned}$$

In the second inequality, we use the following relation:

$$\begin{aligned}
& \mathbb{E}_{k_2}\left[\|\nabla g(y^k) - \tilde{\nabla}g(y^k)\|^2\right] = \mathbb{E}_{k_2}\left[\|\nabla g_{\zeta_k}(\bar{w}^k) - \nabla g_{\zeta_k}(y^k) - \left(\nabla g(\bar{w}^k) - \nabla g(y^k)\right)\|^2\right] \\
\leq\ & \mathbb{E}_{k_2}\left[\|\nabla g_{\zeta_k}(\bar{w}^k) - \nabla g_{\zeta_k}(y^k)\|^2\right] = \sum_{i=1}^{m_2}\frac{1}{\pi_i}\|\nabla g_i(\bar{w}^k) - \nabla g_i(y^k)\|^2 \\
\leq\ & \sum_{i=1}^{m_2}\frac{2L_{g(i)}}{\pi_i}\left(g_i(\bar{w}^k) - g_i(y^k) - \langle\nabla g_i(y^k), \bar{w}^k - y^k\rangle\right) \\
=\ & 2L_g\sum_{i=1}^{m_2}\left(g_i(\bar{w}^k) - g_i(y^k) - \langle\nabla g_i(y^k), \bar{w}^k - y^k\rangle\right) \\
=\ & 2L_g\left(g(\bar{w}^k) - g(y^k) - \langle\nabla g(y^k), \bar{w}^k - y^k\rangle\right),
\end{aligned}$$

(5.7.57)

where the first inequality is from $\mathbb{E}\|\zeta - \mathbb{E}\zeta\|^2 = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}\zeta\|^2 \leq \mathbb{E}\|\zeta\|^2$ and the second inequality is from Theorem 2.1.5 in [66]

Combine (5.7.55), (5.7.56) and (5.7.56):

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[g(v^{k+1})\right] \\
\leq\ & \mathbb{E}_{k_2}\left[(1-\alpha-\beta)g(v^k)+\alpha g(x)+\beta g(\bar{w}^k)\right] \\
& +\mathbb{E}_{k_2}\left[\alpha\langle\tilde{\nabla}g'(y^k), x^{k+0.5}-x\rangle\right]+\left(\frac{\alpha^2 L_g}{2}+\frac{\alpha^2 L_g}{2\beta}+\frac{\alpha\mu_h}{8}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& +\frac{\alpha\mu_h}{8}\mathbb{E}_{k_2}\left[\|x^k-x\|^2\right]+\frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}]+\frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right],
\end{aligned}
$$

implying

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[g(v^{k+1})-g(x)\right] \\
\leq\ & \mathbb{E}_{k_2}\left[(1-\alpha-\beta)\left(g(v^k)-g(x)\right)+\beta\left(g(\bar{w}^k)-g(x)\right)\right] \\
& +\mathbb{E}_{k_2}\left[\alpha\langle\tilde{\nabla}g'(y^k), x^{k+0.5}-x\rangle\right]+\left(\frac{\alpha^2 L_g}{2}+\frac{\alpha^2 L_g}{2\beta}+\frac{\alpha\mu_h}{8}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& +\frac{\alpha\mu_h}{8}\mathbb{E}_{k_2}\left[\|x^k-x\|^2\right]+\frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}]+\frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right],
\end{aligned}
$$

completing the proof.

5.7.3 Proof of Theorem 5.2.3

In view of Lemma 5.2.1 and Lemma 5.2.2, we can establish the following inequality

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[Q(v^{k+1};x)\right] = \mathbb{E}_{k_2}\left[\langle H(x), v^{k+1}-x\rangle + g(v^{k+1}) - g(x)\right] \\
= \; & \mathbb{E}_{k_2}\left[(1-\alpha-\beta)\langle H(x), v^k - x\rangle + \alpha\langle H(x), x^{k+0.5}-x\rangle + \beta\langle H(x), \bar{w}^k - x\rangle\right] \\
& + \mathbb{E}_{k_2}\left[g(v^{k+1}) - g(x)\right] \\
\leq \; & \mathbb{E}_{k_2}\left[(1-\alpha-\beta)\langle H(x), v^k - x\rangle + \alpha\langle H(x), x^{k+0.5}-x\rangle + \beta\langle H(x), \bar{w}^k - x\rangle\right] \\
& + \mathbb{E}_{k_2}\left[(1-\alpha-\beta)\left(g(v^k) - g(x)\right) + \beta\left(g(\bar{w}^k) - g(x)\right)\right] \\
& + \mathbb{E}_{k_2}\left[\alpha\langle \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] + \left(\frac{\alpha^2 L_g}{2} + \frac{\alpha^2 L_g}{2\beta} + \frac{\alpha\mu_h}{8}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5} - x^k\|^2\right] \\
& + \frac{\alpha\mu_h}{8}\mathbb{E}_{k_2}\left[\|x^k - x\|^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k} + \rho_{y^k})^2\right] \\
= \; & (1-\alpha-\beta)\mathbb{E}_{k_2}\left[\langle H(x), v^k - x\rangle + g(v^k) - g(x)\right] + \beta\mathbb{E}_{k_2}\left[\langle H(x), \bar{w}^k - x\rangle + g(\bar{w}^k) - g(x)\right] \\
& + \alpha\mathbb{E}_{k_2}\left[\langle H(x) + \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle + \left(\frac{\alpha L_g}{2} + \frac{\alpha L_g}{2\beta} + \frac{\mu_h}{8}\right)\|x^{k+0.5} - \bar{x}^k\|^2\right] \\
& + \frac{\alpha\mu_h}{8}\mathbb{E}_{k_2}\left[\|x^k - x\|^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k} + \rho_{y^k})^2\right] \\
\leq \; & (1-\alpha-\beta)\mathbb{E}_{k_2}\left[Q(v^k;x)\right] + \beta\mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right] \\
& + \frac{\alpha}{2\gamma}\mathbb{E}_{k_2}\left[\left(1 - p_1 - \frac{1}{2}\gamma\mu_h + \frac{\gamma\mu_h}{4}\right)\|x^k - x\|^2 + p_1\|w^k - x\|^2 - \|x^{k+1} - x\|^2\right] \\
& - \frac{\alpha}{2\gamma}\left(p_1 - 2\gamma^2 L_h^2\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5} - w^k\|^2\right] \\
& - \frac{\alpha}{2\gamma}\left(1 - p_1 - \gamma\mu_h - \alpha\gamma L_g - \frac{\alpha\gamma L_g}{\beta} - \frac{\gamma\mu_h}{4}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5} - x^k\|^2\right] \\
& + \frac{\alpha\mathbb{E}_{k_2}\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]}{2\mu_h} + \alpha\gamma\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}} + \varepsilon_{w^k})^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k} + \rho_{y^k})^2\right],
\end{aligned}
$$

where the last inequality is due to Lemma 5.2.1.

Our goal now is to construct a proper potential function while keeping the coefficients of $\|x^{k+0.5} - w^k\|^2$ and $\|x^{k+0.5} - x^k\|^2$ non-positive. To this end, let us introduce the following bound while noting the expectation $\mathbb{E}_{k_1+}[\cdot] := \mathbb{E}[\cdot|w^k, x^{k+1}]$:

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[\|w^{k+1} - x\|^2\right] \\
= \; & \mathbb{E}_{k_2}\left[\mathbb{E}_{k_1+}\left[\|w^{k+1} - x\|^2\right]\right] = p_1\mathbb{E}_{k_2}\left[\|x^{k+1} - x\|^2\right] + (1-p_1)\mathbb{E}_{k_2}\left[\|w^k - x\|^2\right] \\
= \; & \mathbb{E}_{k_2}\left[p_1\|x^{k+1} - x\|^2 + (1-p_1-c)\|w^k - x\|^2 + c\|w^k - x\|^2\right] \\
\leq \; & \mathbb{E}_{k_2}\left[p_1\|x^{k+1} - x\|^2 + (1-p_1-c)\|w^k - x\|^2\right] \\
& + \mathbb{E}_{k_2}\left[2c\|x^k - x\|^2 + 4c\|x^k - x^{k+0.5}\|^2 + 4c\|x^{k+0.5} - w^k\|^2\right],
\end{aligned}
$$

where $c > 0$ is a parameter that needs to satisfy certain constraints to be determined later. Combine the above inequality with the previous inequality on $Q(v^{k+1};x)$, we have:

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[Q(v^{k+1};x)\right] + \frac{\alpha}{2\gamma}\mathbb{E}_{k_2}\left[(1-p_1)\|x^{k+1}-x\|^2 + \|w^{k+1}-x\|^2\right] \\
\leq\ & (1-\alpha-\beta)\mathbb{E}_{k_2}\left[Q(v^k;x)\right] + \beta\mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right] \\
& + \frac{\alpha}{2\gamma}\mathbb{E}_{k_2}\left[(1-p_1-\frac{1}{4}\gamma\mu_h+2c)\|x^k-x\|^2 + (1-c)\|w^k-x\|^2\right] \\
& - \frac{\alpha}{2\gamma}\left(p_1 - 2\gamma^2 L_h^2 - 4c\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-w^k\|^2\right] \\
& - \frac{\alpha}{2\gamma}\left(1-p_1-\frac{5}{4}\gamma\mu_h - \alpha\gamma L_g - \frac{\alpha\gamma L_g}{\beta} - 4c\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
(5.7.58) & + \frac{\alpha\mathbb{E}_{k_2}\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]}{2\mu_h} + \alpha\gamma\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right].
\end{aligned}
$$

By imposing the following constraints on c:

$$
\begin{cases} 1-\frac{1}{4}\gamma\mu_h + 2c \leq 1, \\ 1-\frac{1}{4}\gamma\mu_h + 2c \geq 1-c \end{cases} \iff \frac{\gamma\mu_h}{12} \leq c \leq \frac{\gamma\mu_h}{8},
$$

let us first take $c = \frac{\gamma\mu_h}{12}$ and impose another set of constraints on γ and α:

$$
(5.7.59) \qquad \begin{cases} p_1 - 2\gamma^2 L_h^2 - \frac{\gamma\mu_h}{3} \geq 0, \\ 1-p_1-\frac{19\gamma\mu_h}{12} - \alpha\gamma L_g - \frac{\alpha\gamma L_g}{\beta} \geq 0, \end{cases}
$$

then (5.7.58) can be reduced to:

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[Q(v^{k+1};x)\right] + \frac{\alpha}{2\gamma}\mathbb{E}_{k_2}\left[(1-p_1)\|x^{k+1}-x\|^2 + \|w^{k+1}-x\|^2\right] \\
\leq\ & (1-\alpha-\beta)\mathbb{E}_{k_2}\left[Q(v^k;x)\right] + \beta\mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right] + \frac{\alpha E_{k_2}\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]}{2\mu_h} + \alpha\gamma\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] \\
& + \frac{\alpha}{2\gamma}\mathbb{E}_{k_2}\left[\left(1-p_1-\frac{\gamma\mu_h}{12}\right)\|x^k-x\|^2 + \left(1-\frac{\gamma\mu_h}{12}\right)\|w^k-x\|^2\right] \\
& + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right] \\
\leq\ & (1-\alpha-\beta)\mathbb{E}_{k_2}\left[Q(v^k;x)\right] + \beta\mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right] + \frac{\alpha\mathbb{E}_{k_2}\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]}{2\mu_h} + \alpha\gamma\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] \\
& + \left(1-\frac{\gamma\mu_h}{12}\right)\frac{\alpha}{2\gamma}\mathbb{E}_{k_2}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] \\
& + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right].
\end{aligned}
$$

Now we only need to add the term $Q(\bar{w}^{k+1};x)$ to the LHS, by noting:

$$
\begin{aligned}
\phi\mathbb{E}_{k_2}\left[Q(\bar{w}^{k+1};x)\right] &= \phi\mathbb{E}_{k_2}\left[\mathbb{E}_{k_2+}\left[Q(\bar{w}^{k+1};x)\right]\right] \\
&= \phi p_2\mathbb{E}_{k_2}\left[Q(v^{k+1};x)\right] + \phi(1-p_2)\mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right],
\end{aligned}
$$

for any $\phi > 0$.

Add the above identity to the previous inequality and take the total expectation, we have:

$$\begin{aligned}
& \mathbb{E}\left[(1-\phi p_2)Q(v^{k+1};x) + \phi Q(\bar{w}^{k+1};x)\right] + \frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^{k+1}-x\|^2 + \|w^{k+1}-x\|^2\right] \\
\leq\ & \mathbb{E}\left[(1-\alpha-\beta)Q(v^k;x) + (\beta+\phi(1-p_2))Q(\bar{w}^k;x)\right] \\
& + \left(1-\frac{\gamma\mu_h}{12}\right)\frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] \\
& + \mathbb{E}\left[\frac{\alpha\mathbb{E}_{k_2}\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]}{2\mu_h} + \alpha\gamma\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}] + \frac{8\alpha}{\mu_h}\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right]\right],
\end{aligned} \tag{5.7.60}$$

where

$$\begin{cases}
\mathbb{E}\left[\mathbb{E}_{k_2}\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]\right] = \mathbb{E}\left[\mathbb{E}'\left[\bar{\varepsilon}^2_{x^{k+0.5}}\right]\right] \leq 2m_1\sigma_h^2 + 2m_1^2\delta_h^2 \\
\mathbb{E}\left[\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right]\right] = \mathbb{E}\left[\mathbb{E}'\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right]\right] \\
\leq 2\cdot\left(2L_h\cdot(\sigma_h^2+\delta_h^2)\cdot\sum_{i=1}^{m_1}\frac{1}{L_{h(i)}}\right) = 2\tilde{\sigma}_h^2.
\end{cases}$$

$$\begin{cases}
\mathbb{E}\left[\mathbb{E}_{k_2}[\bar{\rho}^2_{\bar{w}^k}]\right] = \mathbb{E}\left[\mathbb{E}'[\bar{\rho}^2_{\bar{w}^k}]\right] \leq 2m_2\sigma_g^2 + 2m_2^2\delta_g^2 \\
\mathbb{E}\left[\mathbb{E}_{k_2}\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right]\right] = \mathbb{E}\left[\mathbb{E}'\left[(\rho_{\bar{w}^k}+\rho_{y^k})^2\right]\right] \leq 2\cdot\left(2L_g\cdot(\sigma_g^2+\delta_g^2)\cdot\sum_{i=1}^{m_2}\frac{1}{L_{g(i)}}\right) = 2\tilde{\sigma}_g^2.
\end{cases}$$

By taking $x = x^*$ in (5.7.60) together with the expression of Δ_h and Δ_g, we obtain (5.2.19), thus complete the proof.

5.7.4 Proof of Proposition 5.2.4

We first show that the parameters specified in Proposition 5.2.4 satisfy the constraint (5.7.59). Indeed, the constraints will be reduced to the following:

$$p_1 - \frac{p_1}{8} - \frac{p_1}{12} \geq 0, \quad \frac{15}{16} - \frac{67p_1}{48} \geq 0,$$

where the second inequality holds because $p_1 = \frac{1}{m_1}$ and we assume trivially that $m_1 \geq 2$.

Next, inequality (5.2.19) in Theorem 5.2.3 implies that the reduction rate is given by:

$$C_{\text{red1}} := \max\left\{\frac{1-\alpha-\beta}{1-\phi p_2}, \frac{\beta+\phi(1-p_2)}{\phi}, 1-\frac{\gamma\mu_h}{12}\right\}.$$

With the choice of ϕ and p_2, the following bounds hold:

$$\frac{1-\alpha-\beta}{1-\phi p_2} = \frac{1-2\alpha}{1-\alpha} \leq 1-\alpha,$$

and

$$\frac{\beta+\phi(1-p_2)}{\phi} = 1 - \frac{1}{m_2} + \frac{1}{(1+\alpha)m_2} = 1 - \frac{\alpha}{(1+\alpha)m_2} \leq 1 - \frac{\alpha}{2m_2}.$$

Therefore, the reduction rate is can be further expressed as

$$
\begin{aligned}
& \max\left\{\frac{1-\alpha-\beta}{1-\phi p_2}, \frac{\beta+\phi(1-p_2)}{\phi}, 1-\frac{\gamma\mu_h}{12}\right\} \leq \max\left\{1-\alpha, 1-\frac{\alpha}{2m_2}, 1-\frac{\gamma\mu_h}{12}\right\} \\
= \quad & \max\left\{1-\frac{\alpha}{2m_2}, 1-\frac{\gamma\mu_h}{12}\right\} \\
= \quad & \max\left\{\max\left(1-\frac{\sqrt{\mu_h}}{24\sqrt{L_g m_2}}, 1-\frac{1}{24m_2}\right), \max\left(1-\frac{\mu_h}{48L_h\sqrt{m_1}}, 1-\frac{\sqrt{\mu_h}}{48\sqrt{L_g m_2}}, 1-\frac{1}{48m_1}\right)\right\} \\
= \quad & \max\left\{1-\frac{\sqrt{\mu_h}}{24\sqrt{L_g m_2}}, 1-\frac{1}{24m_2}, 1-\frac{\mu_h}{48L_h\sqrt{m_1}}, 1-\frac{\sqrt{\mu_h}}{48\sqrt{L_g m_2}}, 1-\frac{1}{48m_1}\right\} := C_{\text{red2}},
\end{aligned}
$$

and (5.2.19) becomes

$$
\begin{aligned}
& \mathbb{E}\left[(1-\phi p_2)Q(v^{k+1};x^*) + \phi Q(\bar{w}^{k+1};x^*)\right] + \frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^{k+1}-x^*\|^2 + \|w^{k+1}-x^*\|^2\right] \\
\leq \quad & \mathbb{E}\left[(1-\alpha-\beta)Q(v^k;x^*) + (\beta+\phi(1-p_2))Q(\bar{w}^k;x^*)\right] \\
& + \left(1-\frac{\gamma\mu_h}{12}\right)\frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^k-x^*\|^2 + \|w^k-x^*\|^2\right] + \Delta_h + \Delta_g \\
\leq \quad & C_{\text{red1}} \cdot \left(\mathbb{E}\left[(1-\phi p_2)Q(v^k;x^*) + \phi Q(\bar{w}^k;x^*)\right] + \frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^k-x^*\|^2 + \|w^k-x^*\|^2\right]\right) \\
& + \Delta_h + \Delta_g \\
\leq \quad & C_{\text{red2}} \cdot \left(\mathbb{E}\left[(1-\phi p_2)Q(v^k;x^*) + \phi Q(\bar{w}^k;x^*)\right] + \frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^k-x^*\|^2 + \|w^k-x^*\|^2\right]\right) \\
& + \Delta_h + \Delta_g \\
\leq \quad & C_{\text{red2}}^{k+1} \cdot \left(\mathbb{E}\left[(1-\phi p_2)Q(v^0;x^*) + \phi Q(\bar{w}^0;x^*)\right] + \frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^0-x^*\|^2 + \|w^0-x^*\|^2\right]\right) \\
& + \sum_{i=0}^{k} C_{\text{red2}}^i \left(\Delta_h + \Delta_g\right).
\end{aligned}
$$

Note $v^0 := \bar{w}^0 := w^0 = x^0$. Therefore,

$$
\begin{aligned}
& \mathbb{E}\left[(1-p_1)\|x^{k+1}-x^*\|^2 + \|w^{k+1}-x^*\|^2\right] \\
\leq \quad & \frac{2\gamma}{\alpha} C_{\text{red2}}^{k+1} \left(\mathbb{E}\left[(1-\phi p_2)Q(v^0;x^*) + \phi Q(\bar{w}^0;x^*)\right] + \frac{\alpha}{2\gamma}\mathbb{E}\left[(1-p_1)\|x^0-x^*\|^2 + \|w^0-x^*\|^2\right]\right) \\
& + \frac{2\gamma}{\alpha}\sum_{i=0}^{k} C_{\text{red2}}^i \left(\Delta_h + \Delta_g\right) \\
\leq \quad & C_{\text{red2}}^{k+1} \cdot \left(\frac{4\gamma}{\alpha} Q(x^0;x^*) + 2\|x^0-x^*\|^2\right) + \frac{2\gamma}{\alpha}\sum_{i=0}^{k} C_{\text{red2}}^i \left(\Delta_h + \Delta_g\right) \\
\leq \quad & C_{\text{red2}}^{k+1} \cdot \left(\frac{\gamma}{\alpha\mu_h}\left\|H(x^0) + \nabla g(x^0)\right\|^2 + 2\|x^0-x^*\|^2\right) + \frac{2\gamma}{\alpha}\sum_{i=0}^{k} C_{\text{red2}}^i \left(\Delta_h + \Delta_g\right)
\end{aligned}
$$

By using the expression of C_{red2}, the above rate guarantees the iteration complexity for

reducing the deterministic error to ϵ is

$$\mathcal{O}\left(\frac{1}{1-C_{\text{red2}}}\log\frac{d_0}{\epsilon}\right) = \mathcal{O}\left(\left(m_1+m_2+\sqrt{\frac{L_g m_2}{\mu_h}}+\frac{L_h\sqrt{m_1}}{\mu_h}\right)\ln\frac{d_0}{\epsilon}\right) \tag{5.7.61}$$

where the expected per iteration gradient cost is $\mathcal{O}(p_1 m_1 + p_2 m_2 + 4) = \mathcal{O}(1)$.

The additional stochastic error (per iteration) has the order:

$$\Delta_h = \mathcal{O}\left(\frac{\sqrt{m_2}(m_1\sigma_h^2 + m_1^2\delta_h^2)}{\sqrt{L_g\mu_h}} + \frac{L_h}{L_g}\cdot(\sigma_h^2+\delta_h^2)\cdot\sum_{i=1}^{m_1}\frac{1}{L_{h(i)}}\right),$$
$$\Delta_g = \mathcal{O}\left(\left(\frac{\sqrt{m_2}}{\sqrt{L_g\mu_h}}\right)\left(m_2\sigma_g^2 + m_2^2\delta_g^2 + L_g\cdot(\sigma_g^2+\delta_g^2)\cdot\sum_{i=1}^{m_2}\frac{1}{L_{g(i)}}\right)\right).$$

The total stochastic error after reducing the deterministic error to ϵ is then multiplied by the factor $\sum_{i=0}^{k} C_{\text{red2}}^i = \mathcal{O}\left(\frac{1}{1-C_{\text{red2}}}\right)$, and is summarized as

$$\mathcal{O}\left(\left(m_1+m_2+\sqrt{\frac{L_g m_2}{\mu_h}}+\frac{L_h\sqrt{m_1}}{\mu_h}\right)\cdot\frac{\gamma}{\alpha}\cdot(\Delta_h+\Delta_g)\right).$$

5.7.5 Proof of Lemma 5.3.3

The proof of this lemma follows the similar logic to the proof of SAVREP in Section 5.2. We first consider the sequences related to the VI mapping: $\{\bar{x}^k\}$, $\{x^{k+0.5}\}$, $\{x^k\}$, $\{w^k\}$. It is immediate that we reach the following inequality:

$$\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma_k\langle\hat{H}'(x^{k+0.5})+\tilde{\nabla}g'(y^k), x-x^{k+1}\rangle\right] \\
\leq\ & \mathbb{E}_{k_1}\left[\gamma_k\langle H(x^{k+0.5})+\tilde{\nabla}g'(y^k), x-x^{k+0.5}\rangle\right] \\
& +\mathbb{E}_{k_1}\left[\gamma_k\langle\hat{H}'(x^{k+0.5})-H(x^{k+0.5}), x-x^{k+0.5}\rangle\right] \\
& +\frac{1}{2}\mathbb{E}_{k_1}\left[\gamma_k^2\|H'_{\xi_k}(x^{k+0.5})-H'_{\xi_k}(w^k)\|^2+(1-p_1)\|x^{k+1}-x^k\|^2\right. \\
& \left.-(1-p_1)\|x^{k+0.5}-x^k\|^2+p_1\|x^{k+1}-w^k\|^2-p_1\|x^{k+0.5}-w^k\|^2\right],
\end{aligned} \tag{5.7.62}$$

which is the same as (5.7.51) in the proof of Lemma 5.2.1, except that the parameter γ_k now depends on the iteration k.

Combining with the bound in (5.7.49), we have:

$$
\begin{aligned}
& \frac{1}{2}\mathbb{E}_{k_1}\left[\|x^{k+1}-x\|^2-(1-p_1)\|x^k-x\|^2-p_1\|w^k-x\|^2+(1-p_1)\|x^{k+0.5}-x^k\|^2\right] \\
& +\mathbb{E}_{k_1}\left[\gamma_k\langle H(x^{k+0.5})+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[\gamma_k^2\|H'_{\xi_k}(x^{k+0.5})-H'_{\xi_k}(w^k)\|^2-p_1\|x^{k+0.5}-w^k\|^2\right] \\
& +\mathbb{E}_{k_1}\left[\gamma_k\langle\hat{H}'(x^{k+0.5})-H(x^{k+0.5}),x-x^{k+0.5}\rangle\right] \\
\leq\ & \frac{1}{2}(2\gamma_k^2L_h^2-p_1)\|x^{k+0.5}-w^k\|^2+\gamma_k^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] \\
& +\mathbb{E}_{k_1}\left[\gamma_k\langle\hat{H}'(x^{k+0.5})-H(x^{k+0.5}),x-x^{k+0.5}\rangle\right],
\end{aligned}
$$

where the last inequality is due to the bound (5.7.53). The monotonicity of $H(\cdot)$ implies:

$$
\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma_k\langle H(x)+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle+\frac{1}{2}(p_1-2\gamma_k^2L_h^2)\|x^{k+0.5}-w^k\|^2\right] \\
\leq\ & \mathbb{E}_{k_1}\left[\gamma_k\langle H(x^{k+0.5})+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle+\frac{1}{2}(p_1-2\gamma_k^2L_h^2)\|x^{k+0.5}-w^k\|^2\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[-\|x^{k+1}-x\|^2+(1-p_1)\|x^k-x\|^2+p_1\|w^k-x\|^2-(1-p_1)\|x^{k+0.5}-x^k\|^2\right] \\
& +\gamma_k^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right]+\mathbb{E}_{k_1}\left[\gamma_k\langle\hat{H}'(x^{k+0.5})-H(x^{k+0.5}),x-x^{k+0.5}\rangle\right].
\end{aligned}
$$

Rearranging the terms, we get:

$$
\begin{aligned}
& \mathbb{E}_{k_1}\left[\gamma_k\langle H(x)+\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle\right] \\
\leq\ & \frac{1}{2}\mathbb{E}_{k_1}\left[(1-p_1)\|x^k-x\|^2+p_1\|w^k-x\|^2-\|x^{k+1}-x\|^2\right] \\
& -\mathbb{E}_{k_1}\left[\frac{1}{2}(p_1-2\gamma_k^2L_h^2)\|x^{k+0.5}-w^k\|^2\right]-\frac{1}{2}(1-p_1)\mathbb{E}_{k_1}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& +\gamma_k^2\mathbb{E}_{k_1}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right]+\mathbb{E}_{k_1}\left[\gamma_k\langle\hat{H}'(x^{k+0.5})-H(x^{k+0.5}),x-x^{k+0.5}\rangle\right]. \qquad (5.7.63)
\end{aligned}
$$

The next part of the analysis follows the similar logic to the proof of Lemma 5.2.2, which establishes the relation among the sequences $\{y^k\}$, $\{v^k\}$, $\{\bar{w}^k\}$. It is immediate that we get an inequality similar to (5.7.55):

$$
\begin{aligned}
g(v^{k+1}) \leq\ & (1-\alpha_k-\beta_k)g(v^k)+\beta_k\left(g(y^k)+\langle\nabla g(y^k),\bar{w}^k-y^k\rangle\right)+\frac{L_g\alpha_k^2}{2}\|x^{k+0.5}-x^k\|^2 \\
& +\alpha_k\left(g(x)+\langle\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle+\langle\nabla g(y^k)-\tilde{\nabla}g(y^k),x^{k+0.5}-x\rangle\right) \\
& +\alpha_k\langle\tilde{\nabla}g(y^k)-\tilde{\nabla}g'(y^k),x^{k+0.5}-x\rangle
\end{aligned}
$$

with parameters α_k, β_k depending on the iterations k. Combined with (5.7.56), we get:

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[g(v^{k+1}) - g(x)\right] \\
\leq\ & \mathbb{E}_{k_2}\left[(1-\alpha_k-\beta_k)\left(g(v^k)-g(x)\right) + \beta_k\left(g(\bar{w}^k)-g(x)\right)\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] + \left(\frac{\alpha_k^2 L_g}{2} + \frac{\alpha_k^2 L_g}{2\beta_k}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g(y^k) - \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right].
\end{aligned}
$$

The next steps follow similarly the proof of Theorem 5.2.3, by noticing:

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[Q(v^{k+1};x)\right] \\
=\ & \mathbb{E}_{k_2}\left[\langle H(x), v^{k+1}-x\rangle + g(v^{k+1}) - g(x)\right] \\
=\ & \mathbb{E}_{k_2}\left[(1-\alpha_k-\beta_k)\langle H(x), v^k - x\rangle + \alpha_k\langle H(x), x^{k+0.5}-x\rangle + \beta_k\langle H(x), \bar{w}^k - x\rangle\right] \\
& +\mathbb{E}_{k_2}\left[g(v^{k+1}) - g(x)\right] \\
\leq\ & \mathbb{E}_{k_2}\left[(1-\alpha_k-\beta_k)\langle H(x), v^k - x\rangle + \alpha_k\langle H(x), x^{k+0.5}-x\rangle + \beta_k\langle H(x), \bar{w}^k - x\rangle\right] \\
& +\mathbb{E}_{k_2}\left[(1-\alpha_k-\beta_k)\left(g(v^k)-g(x)\right) + \beta_k\left(g(\bar{w}^k)-g(x)\right)\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] + \left(\frac{\alpha_k^2 L_g}{2} + \frac{\alpha_k^2 L_g}{2\beta_k}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g(y^k) - \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] \\
=\ & (1-\alpha_k-\beta_k)\mathbb{E}_{k_2}\left[\langle H(x), v^k - x\rangle + g(v^k) - g(x)\right] \\
& +\beta_k \mathbb{E}_{k_2}\left[\langle H(x), \bar{w}^k - x\rangle + g(\bar{w}^k) - g(x)\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle H(x) + \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle + \left(\frac{\alpha_k L_g}{2} + \frac{\alpha_k L_g}{2\beta_k}\right)\|x^{k+0.5} - \bar{x}^k\|^2\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g(y^k) - \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right] \\
\overset{(5.7.63)}{\leq}\ & (1-\alpha_k-\beta_k)\mathbb{E}_{k_2}\left[Q(v^k;x)\right] + \beta_k \mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right] \\
& +\frac{\alpha_k}{2\gamma_k}\mathbb{E}_{k_2}\left[(1-p_1)\|x^k - x\|^2 + p_1\|w^k - x\|^2 - \|x^{k+1} - x\|^2\right] \\
& -\frac{\alpha_k}{2\gamma_k}\left(p_1 - 2\gamma_k^2 L_h^2\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5} - w^k\|^2\right] \\
& -\frac{\alpha_k}{2\gamma_k}\left(1 - p_1 - \alpha_k\gamma_k L_g - \frac{\alpha_k\gamma_k L_g}{\beta_k}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5} - x^k\|^2\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right] + \alpha_k\gamma_k \mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}} + \varepsilon_{w^k})^2\right] \\
& +\alpha_k \mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g(y^k) - \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right].
\end{aligned}
$$

Using the relation

$$
\begin{aligned}
\mathbb{E}_{k_2}\left[\|w^{k+1} - x\|^2\right] &= \mathbb{E}_{k_2}\left[\mathbb{E}_{k_1+}\left[\|w^{k+1} - x\|^2\right]\right] \\
&= p_1 \mathbb{E}_{k_2}\left[\|x^{k+1} - x\|^2\right] + (1-p_1)\mathbb{E}_{k_2}\left[\|w^k - x\|^2\right]
\end{aligned}
$$

in the above inequality and rearranging terms, we get the next inequality,

$$
\begin{aligned}
& \mathbb{E}_{k_2}\left[Q(v^{k+1};x)\right] + \frac{\alpha_k}{2\gamma_k}\mathbb{E}_{k_2}\left[(1-p_1)\|x^{k+1}-x\|^2 + \|w^{k+1}-x\|^2\right] \\
\leq\ & (1-\alpha_k-\beta_k)\mathbb{E}_{k_2}\left[Q(v^k;x)\right] + \beta_k\mathbb{E}_{k_2}\left[Q(\bar{w}^k;x)\right] \\
& + \frac{\alpha_k}{2\gamma_k}\mathbb{E}_{k_2}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] \\
& - \frac{\alpha_k}{2\gamma_k}\left(p_1 - 2\gamma_k^2 L_h^2\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-w^k\|^2\right] \\
& - \frac{\alpha_k}{2\gamma_k}\left(1 - p_1 - \alpha_k\gamma_k L_g - \frac{\alpha_k\gamma_k L_g}{\beta_k}\right)\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
& + \alpha_k\mathbb{E}_{k_2}\left[\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right] + \alpha_k\gamma_k\mathbb{E}_{k_2}\left[(\varepsilon_{x^{k+0.5}} + \varepsilon_{w^k})^2\right] \\
& + \alpha_k\mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g(y^k) - \tilde{\nabla} g'(y^k), x^{k+0.5} - x\rangle\right].
\end{aligned}
\tag{5.7.64}
$$

Note that with $\mathbb{E}_{k_1}[\cdot]$, $x^{k+0.5}$ is deterministic and $\mathbb{E}'\left[\mathbb{E}_{k_1}\left[\hat{H}'(x^{k+0.5})\right]\right] = \mathbb{E}'\left[H'(x^{k+0.5})\right]$ $= H(x^{k+0.5})$. Therefore,

$$
\begin{aligned}
& \mathbb{E}\left[\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right] \\
=\ & \mathbb{E}\left[\mathbb{E}'\left[\mathbb{E}_{k_1}\left[\langle \hat{H}'(x^{k+0.5}) - H(x^{k+0.5}), x - x^{k+0.5}\rangle\right]\right]\right] = 0.
\end{aligned}
\tag{5.7.65}
$$

Similarly, we apply the above argument to $\mathbb{E}_{k_2}[\cdot]$ and have

$$
\mathbb{E}_{k_2}\left[\langle \tilde{\nabla} g(y^k) - \tilde{\nabla} g'(y^k), x^k - x\rangle\right] = \langle \nabla g(y^k) - \nabla g'(y^k), x^k - x\rangle
$$

which results in

$$
\mathbb{E}\left[\langle \nabla g(y^k) - \nabla g'(y^k), x^k - x\rangle\right] = \mathbb{E}\left[\mathbb{E}'\left[\langle \nabla g(y^k) - \nabla g'(y^k), x^k - x\rangle\right]\right] = 0. \tag{5.7.66}
$$

Combine inequalities (5.7.64), (5.7.65), (5.7.66) together with the condition (5.3.24), we

have

$$
\begin{aligned}
& \mathbb{E}\left[Q(v^{k+1};x)\right] + \frac{\alpha^k}{2\gamma^k}\mathbb{E}\left[(1-p_1)\|x^{k+1}-x\|^2 + \|w^{k+1}-x\|^2\right] \\
\leq\ & \mathbb{E}\left[(1-\alpha^k-\beta^k)Q(v^k;x) + \beta^k Q(\bar{w}^k;x)\right] + \frac{\alpha^k}{2\gamma^k}\mathbb{E}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] \\
& +\alpha^k\gamma^k\mathbb{E}_k\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] + \alpha^k\mathbb{E}_k\left[\langle\tilde{\nabla}g(y^k)-\tilde{\nabla}g'(y^k), x^{k+0.5}-x^k\rangle\right] \\
& -\frac{\alpha^k}{2\gamma^k}(1-q)\,\mathbb{E}_{k_2}\left[\|x^{k+0.5}-x^k\|^2\right] \\
\leq\ & \mathbb{E}\left[(1-\alpha^k-\beta^k)Q(v^k;x) + \beta^k Q(\bar{w}^k;x)\right] + \frac{\alpha^k}{2\gamma^k}\mathbb{E}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] \\
& +\alpha^k\gamma^k\mathbb{E}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] + \frac{\alpha^k\gamma^k}{2(1-q)}\mathbb{E}\left[\|\tilde{\nabla}g(y^k)-\tilde{\nabla}g'(y^k)\|^2\right] \\
\leq\ & \mathbb{E}\left[(1-\alpha^k-\beta^k)Q(v^k;x) + \beta^k Q(\bar{w}^k;x)\right] + \frac{\alpha^k}{2\gamma^k}\mathbb{E}\left[(1-p_1)\|x^k-x\|^2 + \|w^k-x\|^2\right] \\
& +\alpha^k\gamma^k\Delta,
\end{aligned}
$$

for a constant $0 < q < 1$ to be determined later. Note that since $\mathbb{E}\left[(\varepsilon_{x^{k+0.5}}+\varepsilon_{w^k})^2\right] \leq 2\tilde{\sigma}_h^2$ by the bound (5.2.16), and

$$
\begin{aligned}
& \mathbb{E}\left[\|\tilde{\nabla}g(y^k)-\tilde{\nabla}g'(y^k)\|^2\right] \\
\leq\ & 2\mathbb{E}\left[\left\|\nabla g(\bar{w}^k)-\nabla g'(\bar{w}^k)\right\|^2\right] \\
& +4\mathbb{E}\left[\left\|\nabla g_{\zeta_k}(y^k)-\nabla g'_{\zeta_k}(y^k)\right\|^2\right] + 4\mathbb{E}\left[\left\|\nabla g_{\zeta_k}(\bar{w}^k)-\nabla g'_{\zeta_k}(\bar{w}^k)\right\|^2\right] \\
\overset{(5.2.11),(5.2.17)}{\leq}\ & 4m_2\sigma_g^2 + 8\tilde{\sigma}_g^2,
\end{aligned}
$$

we have $\Delta = 2\tilde{\sigma}_h^2 + \frac{1}{(1-q)}(2m_2\sigma_g^2 + 4\tilde{\sigma}_g^2) = O\left(\sigma_h^2 L_h \sum_{i=1}^{m_1}\frac{1}{L_{h(i)}} + \sigma_g^2 L_g \sum_{i=1}^{m_2}\frac{1}{L_{g(i)}}\right)$.

5.7.6 Proof of Corollary 5.4.3

The unbiasedness of $G'_{i,\varphi}(x,u)$ and $H'_{j,s,\varphi}(x,u)$ follow immediately from Lemma 5.4.2. We shall show the variance bound for $G'_{i,\varphi}(x,u)$, and the proof for $H'_{j,s,\varphi}(x,u)$ follows similarly. Note that:

$$
\begin{aligned}
& \mathbb{E}'_u\left[\left\|G'_{i,\varphi}(x,u)\right\|^2\right] = \mathbb{E}'\left[\mathbb{E}_u\left[\left\|G'_{i,\varphi}(x,u)\right\|^2\right]\right] = \mathbb{E}'\left[\mathbb{E}_u\left[\left\|\frac{n}{\varphi}(g'_i(x+\varphi u)-g'_i(x))u\right\|^2\right]\right] \\
\overset{(5.4.42)}{\leq}\ & \mathbb{E}'\left[2n\left\|\nabla g'_i(x)\right\|^2\right] + \frac{\varphi^2 n^2 L_{i,g}^2}{2} \\
=\ & 2n\mathbb{E}'\left[\|\nabla g_i(x)\|^2 + 2\nabla g_i(x)^\top(\nabla g'_i(x)-\nabla g_i(x)) + \|\nabla g'_i(x)-\nabla g_i(x)\|^2\right] + \frac{\varphi^2 n^2 L_{i,g}^2}{2} \\
\overset{(5.4.40)}{\leq}\ & 2n\left(M_{i,g}^2+\varsigma_g^2\right) + \frac{\varphi^2 n^2 L_{i,g}^2}{2}.
\end{aligned}
$$

In the first inequality, we apply the bound in (5.4.42) on the function $g_i'(\cdot)$, which is the stochastic function estimator of $g_i(\cdot)$. Note that we use the same random variable to estimate the function at the two points $x + \varphi u$ and x when calculating the stochastic zeroth-order gradient $G_{i,\varphi}'(x,u)$. Now, since $\mathbb{E}_u'\left[G_{i,\varphi}'(x,u)\right] = \nabla g_{i,\varphi}(x)$, we have:

$$\begin{aligned} & \mathbb{E}_u'\left[\left\|G_{i,\varphi}'(x,u) - \nabla g_{i,\varphi}(x)\right\|^2\right] = \mathbb{E}_u'\left[\left\|G_{i,\varphi}'(x,u)\right\|^2 - \left\|\nabla g_{i,\varphi}(x)\right\|^2\right] \\ \leq \quad & \mathbb{E}_u'\left[\left\|G_{i,\varphi}'(x,u)\right\|^2\right] \leq 2n\left(M_{i,g}^2 + \varsigma_g^2\right) + \frac{\varphi^2 n^2 L_{i,g}^2}{2}. \end{aligned}$$

5.7.7 Proof of Corollary 5.4.4

We derive the bounds corresponding to $H_j'(z)$. The bounds corresponding to $\nabla g_i'(z)$ can be derived similarly with simpler analysis.

With expressions in (5.4.38) and (5.4.45), we have:

$$\begin{aligned} & \left\|H_j(z) - \mathbb{E}'\left[H_j'(z)\right]\right\|^2 = \left\|Jh_j(x)^\top y - \mathbb{E}'\left[H_{j,\varphi}'(x,u)y\right]\right\|^2 + \left\|h_j(x) - \mathbb{E}'\left[h_j'(x)\right]\right\|^2 \\ = \quad & \left\|\sum_{s=1}^{\ell} y_s\left(\nabla h_{j,s}(x) - \mathbb{E}_u'\left[H_{j,s,\varphi}'(x,u)\right]\right)\right\|^2 = \left\|\sum_{s=1}^{\ell} y_s\left(\nabla h_{j,s}(x) - \nabla h_{j,s,\varphi}(x)\right)\right\|^2 \\ \leq \quad & \left(\sum_{s=1}^{\ell} y_s\left\|\nabla h_{j,s}(x) - \nabla h_{j,s,\varphi}(x)\right\|\right)^2 \overset{5.4.41}{\leq} \left(\frac{\varphi n}{2}\sum_{s=1}^{\ell} y_s L_{j,s,h}\right)^2, \end{aligned}$$

where the second equality is due to $\mathbb{E}'\left[h_{j,s}'(x)\right] = h_{j,s}(x)$ in our assumption (5.4.40). Therefore, by denoting $L_{j,h} := (L_{j,1,h}, L_{j,2,h}, ..., L_{j,s,h})^\top$,

$$\left\|H_j(z) - \mathbb{E}'\left[H_j'(z)\right]\right\| \leq \frac{\varphi n}{2}\sum_{s=1}^{\ell} y_s L_{j,s,h} \leq \frac{\varphi n}{2}\|y\| \cdot \|L_{j,h}\| \leq \frac{\varphi n D_y}{2}\sqrt{\sum_{s=1}^{\ell} L_{j,s,h}^2}.$$

The second bound can be derived by the following:

$$
\begin{aligned}
& \mathbb{E}'\left[\left\|H'_j(z) - \mathbb{E}'\left[H'_j(z)\right]\right\|^2\right] \\
= \quad & \mathbb{E}'\left[\left\|H'_{j,\varphi}(x,u)y - \mathbb{E}'_u\left[H'_{j,\varphi}(x,u)y\right]\right\|^2 + \left\|h'_j(x) - \mathbb{E}'\left[h'_j(x)\right]\right\|^2\right] \\
\overset{(5.4.40)}{\leq} \quad & \mathbb{E}'\left[\left\|H'_{j,\varphi}(x,u)y - \mathbb{E}'_u\left[H'_{j,\varphi}(x,u)y\right]\right\|^2\right] + \ell\varpi^2 \\
= \quad & \mathbb{E}'\left[\left\|\sum_{s=1}^{\ell} y_s\left(H'_{j,s,\varphi}(x,u) - \nabla h_{j,s,\varphi}(x,u)\right)\right\|^2\right] + \ell\varpi^2 \\
\leq \quad & \mathbb{E}'\left[\ell \cdot \sum_{s=1}^{\ell} y_s^2 \left\|H'_{j,s,\varphi}(x,u) - \nabla h_{j,s,\varphi}(x,u)\right\|^2\right] + \ell\varpi^2 \\
\overset{(5.4.44)}{\leq} \quad & \ell \cdot \tilde{\varsigma}_h^2 \cdot \sum_{s=1}^{\ell} y_s^2 + \varpi^2 \leq \ell\tilde{\varsigma}_h^2 D_y^2 + \ell\varpi^2.
\end{aligned}
$$

Chapter 6

Beyond Monotone Variational Inequality Problems: Solution Methods and Iteration Complexities

6.1 Introduction

In this chapter, we discuss variational inequality problems without monotonicity from the perspective of convergence of projection-type algorithms. In particular, we identify existing conditions as well as present new conditions that are sufficient to guarantee convergence. The first half of this chapter focuses on the case where a Minty solution exists (also known as *Minty condition*), which is a common assumption in the recent developments for non-monotone VI. The second half explores alternative sufficient conditions that are different from the existing ones such as monotonicity or Minty condition, using an *algorithm-based* approach. Through examples and convergence analysis, we show that these conditions are capable of characterizing different classes of VI problems where the algorithms are guaranteed to converge.

For non-monotone VI, earlier research has developed non-projection-type methods such as the KKT condition based methods and merit function based methods (see [18, 80, 77, 19] and the references therein). However, it is in general difficult to establish iteration complexity for non-projection-type methods for non-monotone VI. In recent years, research on developing algorithms for non-monotone VI has focused on the VI problems where the so-called *Minty*

solutions exist. A Minty solution to VI is a solution x^* where the following inequality is satisfied:

$$\langle F(x), x - x^* \rangle \geq 0$$

for all $x \in \mathcal{X}$. When the constraint set is a close convex set and the operator F is continuous and monotone, all solutions to the VI (if any) are Minty solutions. The existence of Minty solutions turns out to be critical in establishing convergence for the projection-type methods for non-monotone VI, and there have been recent developments of such results; see e.g. [53, 91, 9, 102, 48].

In this chapter, we follow this line of research on the convergence of projection-type methods for non-monotone VI. We start from the common assumption made in the literature, that is, a Minty solution exists. We show that a general extra-gradient-type method, the ARE update proposed in Chapter 3, converges in a guaranteed rate with a similar convergence behavior as Perseus in [53]. In addition, we are interested in the concept of Minty solution itself, especially in the non-monotone setting where a VI solution is not necessarily a Minty solution. Therefore, we investigate implications given by the Minty solutions in different problem classes such as optimization and Nash games. Finally, we explore the possibilities of alternative sufficient conditions for convergence of projection-type methods through an *algorithm-based* approach. Conventionally, algorithms are devised to ensure convergence under a given problem framework, such as monotone VI or VI with Minty solutions, and the convergence behavior is analyzed within the framework. In this paper, we follow an opposite direction by deriving sufficient conditions for convergence based on the algorithms we are interested in. In other words, for a given algorithm, we aim to identify VI with specific structures where the algorithm is guaranteed to converge to a solution. It turns out that this approach makes it possible to characterize structures of VI models that are different from commonly encountered conditions such as the monotonicity or the Minty condition. We present several conditions of this kind and demonstrate examples as well as proving convergence of gradient projection methods and extra-gradient methods under these conditions.

6.2 Non-monotone VI problems with Minty Solution

6.2.1 Definitions and solution concepts

6.2.1.1 VI solutions and Minty solutions

In order to set the background for the discussion in this chapter, let us first formally define the variational inequality problem and its solution set. For a given set $\mathcal{X} \subseteq \mathbb{R}^n$ and a continuous mapping $F : \mathcal{X} \mapsto \mathbb{R}^n$, consider the following VI model, to be denoted by $\mathrm{VI}(F; \mathcal{X})$:

$$\begin{array}{ll} \text{Find} & x^* \in \mathcal{X} \\ \text{such that} & \langle F(x^*), x - x^* \rangle \geq 0, \ \forall x \in \mathcal{X}. \end{array}$$

Let the solution set of the above model be $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$. It is also referred to as the set of *strong solutions*, or simply the *solutions*, to the VI model. The non-emptiness of $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$ can be guaranteed by imposing some assumptions on the basic problem structure.

Assumption 6.2.1. *F is a continuous mapping. $\mathcal{X}$ is non-empty, convex and compact.*

Assumption 6.2.1 ensures that $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{X})) \neq \emptyset$ [19, 23], and we shall make this assumption throughout the chapter. In addition to the (strong) solutions to the VI, there is another important solution concept, which is the so-called *Minty solutions* or *weak solutions*, defined as the set of x^* such that

$$\langle F(x), x - x^* \rangle \geq 0, \ \forall x \in \mathcal{X}.$$

Let the set of Minty solutions to $\mathrm{VI}(F; \mathcal{X})$ be denoted by $\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{X}))$. The well-known Minty's Lemma states the follwing:

Lemma 6.2.2 (Minty's Lemma)**.** *If F is continuous, $\mathcal{X}$ is non-empty, closed and convex, then* $\mathrm{Sol}_m(\mathit{VI}(F; \mathcal{X})) \subseteq \mathrm{Sol}(\mathit{VI}(F; \mathcal{X}))$.

If additionally F is monotone, then $\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{X})) = \mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$. Indeed, for every $x^* \in \mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$, we have:

$$\langle F(x), x - x^* \rangle \geq \langle F(x^*), x - x^* \rangle \geq 0,$$

thus $x^* \in \mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{X}))$. In this chapter, while we always assume the non-emptiness of $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$ (by Assumption 6.2.1), we extend the discussion to the broader class of VI where F is not necessarily monotone. Alternatively, we focus on the Minty solutions and assume $\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{X})) \neq \emptyset$ in this section, and we discuss other conditions in Section 6.3 where no Minty solutions exist.

6.2.1.2 Merit functions

In the context of strongly monotone VI where the solution x^* uniquely exists, it is common to use (squared) distance to the solution $\|x - x^*\|^2$ in the iteration complexity analysis. For VI problems that are merely monotone, there are two other merit functions that are widely used, known as the *gap function* and the *dual gap function*. While monotonicity is not assumed in this chapter, we may still use these two merit functions as the measurement of convergence. In this section we re-introduce them in a fashion that relates them with the two solution concepts, VI solutions $\mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$ and Minty solutions $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X}))$, introduced earlier.

Proposition 6.2.3. *Suppose that $\mathcal{X}$ is compact. It holds that*

$$\mathrm{Sol}(\,VI(F;\mathcal{X})) \neq \emptyset \Longleftrightarrow \min_{y\in\mathcal{X}} \max_{x\in\mathcal{X}} \langle F(y), y - x\rangle = 0.$$

Proof. Observe that for any $y \in \mathcal{X}$, we always have

$$\max_{x\in\mathcal{X}} \langle F(y), y - x\rangle \geq \langle F(y), y - y\rangle = 0.$$

Hence, $\min_{y\in\mathcal{X}} \max_{x\in\mathcal{X}} \langle F(y), y - x\rangle \geq 0$, or $\max_{y\in\mathcal{X}} \min_{x\in\mathcal{X}} \langle F(y), x - y\rangle \leq 0$.

$\Longrightarrow$: Choose any $x^* \in \mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$. We have $\min_{x\in\mathcal{X}} \langle F(x^*), x - x^*\rangle = 0$, implying

$$\max_{y\in\mathcal{X}} \min_{x\in\mathcal{X}} \langle F(y), x - y\rangle = 0.$$

$\Longleftarrow$: Let

$$y^* \in \arg\max_{y\in\mathcal{X}} \left[\min_{x\in\mathcal{X}} \langle F(y), x - y\rangle\right].$$

It follows that $\min_{x\in\mathcal{X}} \langle F(y^*), x - y^*\rangle = 0$, or equivalently put

$$\langle F(y^*), x - y^*\rangle \geq 0, \text{ for any } x \in \mathcal{X}.$$

Hence, $y^* \in \mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$. □

In a similar vein, we have:

Proposition 6.2.4. *Suppose that $\mathcal{X}$ is compact. It holds that*

$$\mathrm{Sol}_m(\,VI(F;\mathcal{X})) \neq \emptyset \Longleftrightarrow \max_{x\in\mathcal{X}} \min_{y\in\mathcal{X}} \langle F(y), y - x\rangle = 0.$$

Proof. First, observe that in general, we have

$$\max_{x\in\mathcal{X}} \min_{y\in\mathcal{X}} \langle F(y), y - x\rangle \leq 0$$

because for any given $x \in \mathcal{X}$, it follows that $\min_{y\in\mathcal{X}}\langle F(y), y-x\rangle \le \langle F(x), x-x\rangle = 0$.

$\Longrightarrow$: Choose any $x^* \in \mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X}))$. Since $\langle F(y), y-x^*\rangle \ge 0$ for all $y \in \mathcal{X}$, we have

$$\min_{y\in\mathcal{X}}\langle F(y), y-x^*\rangle = 0,$$

implying $\max_{x\in\mathcal{X}}\min_{y\in\mathcal{X}}\langle F(y), y-x\rangle = 0$.

$\Longleftarrow$: Let

$$x^* \in \arg\max_{x\in\mathcal{X}}\left[\min_{y\in\mathcal{X}}\langle F(y), y-x\rangle\right].$$

We have $\min_{y\in\mathcal{X}}\langle F(y), y-x^*\rangle = 0$, or equivalently

$$\langle F(y), y-x^*\rangle \ge 0, \text{ for all } y \in \mathcal{X},$$

implying $x^* \in \mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X}))$. $\square$

The above analysis naturally leads to the following notions of merit functions:

$$G(x) := \max_{y\in\mathcal{X}}\langle F(x), x-y\rangle,$$

also known as the *gap function*, and

$$H(x) := \max_{y\in\mathcal{X}}\langle F(y), x-y\rangle,$$

also known as the *dual gap function*.

Based on our analysis, we have:

Proposition 6.2.5.

- *$G(x) \ge 0$ for all $x \in \mathcal{X}$, and $G(x) = 0$ if and only if $x \in \mathrm{Sol}(VI(F;\mathcal{X}))$.*
- *$H(x) \ge 0$ for all $x \in \mathcal{X}$, and $H(x) = 0$ if and only if $x \in \mathrm{Sol}_m(VI(F;\mathcal{X}))$.*

Therefore, we may introduce the following notion of ϵ-solutions.

Definition 6.2.6. *For $\epsilon > 0$, we call x to be an ϵ-VI solution if $G(x) \le \epsilon$; we call x to be an ϵ-Minty solution if $H(x) \le \epsilon$.*

6.2.1.3 Relaxation of monotonicity

We remark that there are several conditions can be made on the structure of F, under which the (pure) monotonicity is relaxed but the connections between $\mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$ and $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X}))$ still exist. While we do not assume most of these conditions, we summarize them below for the benefit of easy referencing. In particular, we only consider the *Minty condition* among others, which simply states $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$.

- Weak Sharpness:

$$\langle F(x^*), x - x^* \rangle \geq \mu \|x - x^*\|^2, \quad \forall x \in \mathcal{X}, \quad x^* \in \mathrm{Sol}(\mathrm{VI}(F, \mathcal{X})),$$

  for some $\mu \geq 0$.

- Pseudo-monotonicity:

$$\langle F(y), x - y \rangle \geq 0 \Longrightarrow \langle F(x), x - y \rangle \geq 0, \quad \forall x, y \in \mathcal{X}.$$

- Strongly pseudo-monotonicity:

$$\langle F(y), x - y \rangle \geq 0 \Longrightarrow \langle F(x), x - y \rangle \geq \mu \|x - y\|^2, \quad \forall x, y \in \mathcal{X},$$

  for some $\mu > 0$.

- Quasi-monotonicity:

$$\langle F(y), x - y \rangle > 0 \Longrightarrow \langle F(x), x - y \rangle \geq 0, \quad \forall x, y \in \mathcal{X}.$$

- Minty's condition: $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$, i.e. there exists $x^* \in \mathcal{X}$ such that

$$\langle F(x), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

- Strong Minty's condition (Generalized monotonicity): there exists $x^* \in \mathcal{X}$ such that

$$\langle F(x), x - x^* \rangle \geq \mu \|x - x^*\|^2, \quad \forall x \in \mathcal{X},$$

  for some $\mu \geq 0$.

A few remarks are in place to specify some implications given by the above conditions.

Remark 6.2.7.

- *If F is pseudo-monotone, then* $\mathrm{Sol}(\mathit{VI}(F;\mathcal{X})) \subset \mathrm{Sol}_m(\mathit{VI}(F;\mathcal{X}))$. *If further F is continuous, $\mathcal{X}$ is nonempty, closed and convex, then* $\mathrm{Sol}(\mathit{VI}(F;\mathcal{X})) = \mathrm{Sol}_m(\mathit{VI}(F;\mathcal{X}))$.

- *If $\mathcal{X}$ is closed and bounded, then* $\mathrm{Sol}_m(VI(F;\mathcal{X})) \neq \emptyset$ *if and only if F is quasi-monotone [24].*

- *Assume* $\mathrm{Sol}(VI(F;\mathcal{X})) \neq \emptyset$*, then the following relations hold:*

$$\textit{monotone} \Longrightarrow \textit{pseudo-monotone} \Longrightarrow \textit{Minty's condition}$$

and

$$\textit{strongly monotone} \Longrightarrow \textit{strongly pseudo-monotone} \Longrightarrow \textit{strong Minty's condition}$$

6.2.2 Convergence of projection-type methods

In this section, we present a solution method of projection type that can be shown to converge to Minty solutions by simply assuming $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$. This method is exactly the Approximation-based Regularized Extra-gradient method, or simply ARE, discussed in Chapter 3. While in the previous section it is proposed originally for solving monotone VI with convergence rate $\mathcal{O}(N^{-\frac{p+1}{2}})$, it turns out that ARE not only solves monotone VI, but also solves non-monotone VI as long as $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$. A similar technique in the analysis has been used in [53] to show a different projection-type method—Perseus—can converge to Minty solutions for non-monotone VI at the same rate as to be developed later in this paper.

For clear referencing, let us summarize the update procedure of ARE here:

$$\text{(6.2.1)} \quad \begin{cases} x^{k+0.5} & := & \mathrm{VI}_{\mathcal{X}}\left(\tilde{F}(x;x^k) + L_p\|x-x^k\|^{p-1}(x-x^k)\right), \\ x^{k+1} & := & \arg\min\limits_{x\in\mathcal{X}} \langle F(x^{k+0.5}), x-x^k\rangle + \frac{L_p\|x^{k+0.5}-x^k\|^{p-1}}{2}\|x-x^k\|^2, \end{cases}$$

for $k = 1, 2, ...$, where L_p is the Lipschitz constant for $\nabla^{p-1}F(x)$ satisfying the condition

$$\|\nabla^{p-1}F(x) - \nabla^{p-1}F(y)\| \leq L_p\|x-y\|, \tag{6.2.2}$$

$\tilde{F}(\cdot;y) : \mathbb{R}^n \mapsto \mathbb{R}^n$ is a general approximation mapping estimated at y satisfying the bound:

$$\|\tilde{F}(x;y) - F(x)\| \leq \tau L_p\|x-y\|^p, \tag{6.2.3}$$

and we use the notation $\mathrm{VI}_{\mathcal{X}}(F)$ to denote solving for a solution in $\mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$ as a subroutine in the update. In the ARE update (6.2.1), the subroutine at iteration k specifically solve the VI model associated with the regularized approximation operator $\tilde{F}(x;x^k) + L_p\|x-x^k\|^{p-1}(x-x^k)$.

The initial steps in the analysis will follow the exact same logic as the proof in Theorem 3.2.1. By the definition of $x^{k+0.5}$, we have

$$\langle \tilde{F}(x^{k+0.5}; x^k) + L_p\|x^{k+0.5} - x^k\|^{p-1}(x^{k+0.5} - x^k), x - x^{k+0.5}\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{6.2.4}$$

Denote $\gamma_k = L_p\|x^{k+0.5} - x^k\|^{p-1}$. Substituting $x = x^{k+1}$ in (6.2.4) we have

$$\begin{aligned} & \langle \tilde{F}(x^{k+0.5}; x^k), x^{k+1} - x^{k+0.5}\rangle \\ \geq \ & \gamma_k \langle x^{k+0.5} - x^k, x^{k+0.5} - x^{k+1}\rangle \\ = \ & \frac{\gamma_k}{2}\left(\|x^{k+0.5} - x^k\|^2 + \|x^{k+1} - x^{k+0.5}\|^2 - \|x^{k+1} - x^k\|^2\right). \end{aligned} \tag{6.2.5}$$

On the other hand, by the optimality condition at x^{k+1} we have

$$\langle F(x^{k+0.5}) + \gamma_k(x^{k+1} - x^k), x - x^{k+1}\rangle \geq 0, \text{ for all } x \in \mathcal{X}.$$

Hence,

$$\begin{aligned} & \langle F(x^{k+0.5}), x - x^{k+1}\rangle \\ \geq \ & \gamma_k \langle x^{k+1} - x^k, x^{k+1} - x\rangle \\ = \ & \frac{\gamma_k}{2}\left(\|x^{k+1} - x\|^2 + \|x^{k+1} - x^k\|^2 - \|x^k - x\|^2\right), \text{ for all } x \in \mathcal{X}. \end{aligned} \tag{6.2.6}$$

Continue with the above inequality, for any given $x \in \mathcal{X}$ we have

$$\begin{aligned} & \frac{\gamma_k}{2}\left(\|x^{k+1} - x\|^2 + \|x^{k+1} - x^k\|^2 - \|x^k - x\|^2\right) \\ \overset{(6.2.6)}{\leq} \ & \langle F(x^{k+0.5}), x - x^{k+1}\rangle \\ = \ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \langle F(x^{k+0.5}), x^{k+0.5} - x^{k+1}\rangle \\ = \ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \langle F(x^{k+0.5}) - \tilde{F}(x^{k+0.5}; x^k), x^{k+0.5} - x^{k+1}\rangle \\ & + \langle \tilde{F}(x^{k+0.5}; x^k), x^{k+0.5} - x^{k+1}\rangle \\ \leq \ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \|F(x^{k+0.5}) - \tilde{F}(x^{k+0.5}; x^k)\| \cdot \|x^{k+0.5} - x^{k+1}\| \\ & + \langle \tilde{F}(x^{k+0.5}; x^k), x^{k+0.5} - x^{k+1}\rangle \\ \leq \ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \frac{\|F(x^{k+0.5}) - \tilde{F}(x^{k+0.5}; x^k)\|^2}{2\gamma_k} + \frac{\gamma_k\|x^{k+0.5} - x^{k+1}\|^2}{2} \\ & + \langle \tilde{F}(x^{k+0.5}; x^k), x^{k+0.5} - x^{k+1}\rangle \\ \overset{(6.2.3)}{\leq} \ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \frac{\tau^2 L_p^2\|x^{k+0.5} - x^k\|^{2p}}{2\gamma_k} + \frac{\gamma_k\|x^{k+0.5} - x^{k+1}\|^2}{2} \\ & + \langle \tilde{F}(x^{k+0.5}; x^k), x^{k+0.5} - x^{k+1}\rangle. \end{aligned}$$

Noticing that $\frac{\tau^2 L_p^2 \|x^{k+0.5}-x^k\|^{2p}}{2\gamma_k} = \frac{\tau^2\gamma_k\|x^{k+0.5}-x^k\|^2}{2}$, and further using (6.2.5) we derive from the above that

$$\begin{aligned}
& \frac{\gamma_k}{2}\left(\|x^{k+1}-x\|^2 + \|x^{k+1}-x^k\|^2 - \|x^k - x\|^2\right) \\
\leq\ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \frac{\tau^2\gamma_k\|x^{k+0.5}-x^k\|^2}{2} + \frac{\gamma_k\|x^{k+0.5}-x^{k+1}\|^2}{2} \\
& + \frac{\gamma_k}{2}\left[-\|x^{k+0.5}-x^k\|^2 - \|x^{k+1}-x^{k+0.5}\|^2 + \|x^{k+1}-x^k\|^2\right].
\end{aligned}$$

Canceling out terms, we simplify the above inequality into

(6.2.7)
$$\langle F(x^{k+0.5}), x^{k+0.5} - x\rangle + \frac{\gamma_k}{2}\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \leq \frac{\gamma_k}{2}\left[\|x^k - x\|^2 - \|x^{k+1}-x\|^2\right].$$

In the original analysis in Theorem 3.2.1, the rest of the proof continues with the monotonicity of F. In this analysis, we assume the Minty condition (i.e. $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$) holds instead of monotonicity of F. Taking any fixed $x = x^* \in \mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X}))$ in the above inequality, we have:

$$\left(1-\tau^2\right)\|x^{k+0.5}-x^k\|^2 \leq \left[\|x^k - x^*\|^2 - \|x^{k+1}-x^*\|^2\right], \tag{6.2.8}$$

since $\langle F(x^{k+0.5}), x^{k+0.5} - x^*\rangle \geq 0$. Summing up this inequality for $k = 1, ..., N$ gives us:

$$\min_{1\leq k\leq N}\|x^{k+0.5}-x^k\|^2 \leq \frac{1}{N}\sum_{k=1}^{N}\|x^{k+0.5}-x^k\|^2 \leq \frac{1}{N(1-\tau^2)}\|x^1 - x^*\|^2. \tag{6.2.9}$$

Condition (6.2.4) for updating $x^{k+0.5}$ implies for all $x \in \mathcal{X}$, we have:

$$\begin{aligned}
\langle \tilde{F}(x^{k+0.5}; x^k), x^{k+0.5} - x\rangle &\leq -L_p\|x^{k+0.5}-x^k\|^{p-1}(x^{k+0.5}-x^k)^\top(x^{k+0.5}-x) \\
&\leq L_p D\|x^{k+0.5}-x^k\|^p,
\end{aligned}$$

where $D := \max_{x,x'\in\mathcal{X}}\|x - x'\|$.

Denote $k_N := \arg\min_{1\leq k\leq N}\|x^{k+0.5}-x^k\|^2$, we have:

$$\begin{aligned}
& \langle F(x^{k_N+0.5}), x^{k_N+0.5} - x\rangle \\
=\ & \langle F(x^{k_N+0.5}) - \tilde{F}(x^{k_N+0.5}; x^{k_N}), x^{k_N+0.5} - x\rangle + \langle \tilde{F}(x^{k_N+0.5}; x^{k_N}), x^{k_N+0.5} - x\rangle \\
\leq\ & \left\|F(x^{k_N+0.5}) - \tilde{F}(x^{k_N+0.5}; x^{k_N})\right\| \cdot \|x^{k_N+0.5} - x\| + L_p D\|x^{k_N+0.5} - x^{k_N}\|^p \\
\leq\ & (1+\tau)L_p D\|x^{k_N+0.5} - x^{k_N}\|^p \\
\leq\ & (1+\tau)L_p D\frac{1}{N^{\frac{p}{2}}(1-\tau^2)^{\frac{p}{2}}}\|x^1 - x^*\|^p \leq \frac{(1+\tau)L_p D^{p+1}}{N^{\frac{p}{2}}(1-\tau^2)^{\frac{p}{2}}},
\end{aligned}$$

which holds for all $x \in \mathcal{X}$. Therefore,

$$G(x^{k_N+0.5}) = \max_{x\in\mathcal{X}} \langle F(x^{k_N+0.5}), x^{k_N+0.5} - x\rangle \leq \frac{(1+\tau)L_p D^{p+1}}{N^{\frac{p}{2}}(1-\tau^2)^{\frac{p}{2}}}.$$

We summarize the above results in the next theorem.

Theorem 6.2.8. *Consider the ARE update (6.2.1). Suppose that conditions (6.2.2) and (6.2.3) are satisfied, and* $\mathrm{Sol}_m(VI(F;\mathcal{X})) \neq \emptyset$. *Then, the sequence produced by ARE converges at the following rate:*

$$\|x^{k_N+0.5} - x^{k_N}\|^2 = \mathcal{O}(1/N), \quad G(x^{k_N+0.5}) = \mathcal{O}(1/N^{\frac{p}{2}}).$$

Remark 6.2.9. *In the above analysis, by using the same* $x^* \in \mathrm{Sol}_m(VI(F;\mathcal{X}))$ *in (6.2.8) repetitively, it can be seen that the sequence* $\{x^k\}$ *converges to the specific Minty solution in terms of squared distance. On the other hand, while* $\|x^{k_N+0.5} - x^{k_N}\|^2$ *also converges at a rate* $1/N$ *by (6.2.9), the final result guarantees a rate of convergence* $1/N^{\frac{p}{2}}$ *in terms of the merit function* $G(x^{k_N+0.5})$*, which gives an* ϵ*-VI solution (but not necessarily an* ϵ*-Minty solution) based on Definition 6.2.6.*

Finally, we note that if F is monotone (in which case $\mathrm{Sol}(\mathrm{VI}(F;\mathcal{X})) = \mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X}))$), the convergence rate is $1/N^{\frac{p+1}{2}}$ in terms of the merit function $H(\bar{x}_N)$, where $\bar{x}_N$ is the weighted average of $x^{k+0.5}$ (see Theorem 3.2.1). It is an improved rate compared to the above result where we only assume $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$.

6.2.3 Minty solutions beyond general VI problems

In the previous section, we see that the existence of Minty solutions is indeed an important property to have for solving non-monotone VI. Without assuming any other conditions, it allows projection-type methods such as extra-gradient method (a first-order specialized method of ARE) to converge (to a Minty solution) with guaranteed rate. It is then natural to ask whether the same solution concept presents with similar significance in other problem classes related to VI and what are the implications of the Minty solution therein. In this section, we first discuss the role of the Minty solution in optimization. The discussion proceeds in the context of Nash games, where we present the implications of the Minty solution and its connections to the VI model.

6.2.3.1 Minty solutions in optimization

Consider the optimization problem:

$$\min_{x \in \mathcal{X}} \; f(x), \tag{6.2.10}$$

where $f(x)$ is continuously differentiable, $\mathcal{X}$ is convex and closed. The local first-order optimality condition is given by:

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X}, \tag{6.2.11}$$

which is equivalent to the VI model $\mathrm{VI}(\nabla f; \mathcal{X})$ with solution set $\mathrm{Sol}(\mathrm{VI}(\nabla f; \mathcal{X}))$. Now suppose a Minty solution exists for this VI model, that is, $\mathrm{Sol}_m(\mathrm{VI}(\nabla f; \mathcal{X})) \neq \emptyset$ and any element $x^* \in \mathrm{Sol}_m(\mathrm{VI}(\nabla f; \mathcal{X}))$ satisfies

$$\langle \nabla f(x), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X}. \tag{6.2.12}$$

Note that Minty's Lemma applies here due to our assumption. Therefore, $\mathrm{Sol}_m(\mathrm{VI}(\nabla f; \mathcal{X})) \subseteq \mathrm{Sol}(\mathrm{VI}(\nabla f; \mathcal{X}))$, and any x^* satisfying (6.2.12) is a local first-order stationary point ((6.2.11) holds). The next Theorem states that a Minty solution in optimization, if exists, is in fact a global solution to the problem.

Theorem 6.2.10 (Optimality of Minty solution)**.** *For the optimization problem* (6.2.10) *where f is continuously differentiable, $\mathcal{X}$ is convex and closed. The following holds:*

$$\begin{aligned} & \langle \nabla f(x), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X} \\ \Longrightarrow \quad & f(x^*) \leq f(x), \quad \forall x \in \mathcal{X}. \end{aligned}$$

In other words, if a Minty solution exists, it is a global solution to the problem.

Proof. *Using the following identity:*

$$f(x^*) - f(x) = \int_{t=0}^{1} \nabla f(x + t(x^* - x))^\top (x^* - x) dt,$$

we have for $t = 1$, $\nabla f(x^*)^\top (x^* - x) \leq 0$ since (6.2.11) *holds. For $t = 0$, we have $\nabla f(x)^\top (x^* - x) \leq 0$ due to* (6.2.12)*. For $0 < t < 1$, let $\hat{x} = x + t(x^* - x) \in \mathcal{X}$, then*

$$\nabla f(x + t(x^* - x))^\top (x^* - x) = \frac{1}{1-t} \nabla f(\hat{x})^\top (x^* - \hat{x}) \leq 0, \quad 0 < t < 1,$$

where the last inequality is again due to (6.2.12)*. Therefore, we can conclude that*

$$\begin{aligned} & f(x^*) - f(x) = \int_{t=0}^{1} \nabla f(x + t(x^* - x))^\top (x^* - x) dt \leq 0 \\ \Longrightarrow \quad & f(x^*) \leq f(x). \end{aligned}$$

□

Remark 6.2.11. *The Minty solution is always a global solution in optimization, provided that $f(x)$ is continuously differentiable and $\mathcal{X}$ is closed convex set. However, a global solution needs not be a Minty solution. For a meaningful optimization problem where a global solution exists, a Minty solution may not exist.*

Consider a one-dimensional optimization problem:

$$\min_{-1\leq x\leq 1} \; -x^2,$$

the global solutions are $x^* = -1, 1$. For $x^* = -1$:

$$\langle \nabla f(x), x - x^* \rangle = \langle -2x, x+1 \rangle < 0, \; 0 < x \leq 1;$$

for $x^* = 1$,

$$\langle \nabla f(x), x - x^* \rangle = \langle -2x, x-1 \rangle < 0, \; -1 \leq x < 0.$$

Therefore, neither of the global solutions is a Minty solution. Same as the VI model, when the objective function is convex, the set of local solutions ($\mathrm{Sol}(\mathrm{VI}(\nabla f; \mathcal{X}))$) coincides with the set of Minty solutions ($\mathrm{Sol}_m(\mathrm{VI}(\nabla f; \mathcal{X}))$), thus every global solution is a Minty solution.

6.2.3.2 Minty solutions in Games

Consider a two-player game:

$$\begin{cases} x: & \min\limits_{x\in\mathcal{X}} \; \theta_x(x,y) \\ y: & \min\limits_{y\in\mathcal{Y}} \; \theta_y(x,y), \end{cases} \tag{6.2.13}$$

where we use x, y to denote both the players and their corresponding strategies. Assume θ_x, θ_y are both continuously differentiable for fixed y, x, and $\mathcal{X}, \mathcal{Y}$ are closed convex sets. Let us first define three different notions of equilibria in this game, starting from the well-known Nash equilibrium.

Definition 6.2.12 (Nash equilibrium (NE))**.** *A solution pair $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is said to be in Nash equilibrium if and only if*

$$\theta_x(x^*, y^*) \leq \theta_x(x, y^*), \; \forall x \in \mathcal{X}, \quad \theta_y(x^*, y^*) \leq \theta_y(x^*, y), \; \forall y \in \mathcal{Y}.$$

In other words, for player x it is not possible to be better off by deviating from the Nash equilibrium strategy x^* if the opponent continues to play $y = y^*$ and vice versa. For a Nash equilibrium pair (x^*, y^*), x^*/y^* is the global minimizer of the objective function $\theta_x(\cdot, y^*)/\theta_y(x^*, \cdot)$ for fixed y^*/x^*.

Definition 6.2.13 (Quasi-Nash equilibrium [75] (QNE))**.** *A solution pair* $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ *is said to be in quasi-Nash equilibrium if and only if*

$$\langle \nabla_x \theta_x(x^*, y^*), x - x^* \rangle \geq 0, \forall x \in \mathcal{X}, \quad \langle \nabla_y \theta_y(x^*, y^*), y - y^* \rangle \geq 0, \forall y \in \mathcal{Y}.$$

Unlike Nash equilibrium where x^* and y^* have to be global minimizers of their respective objective functions when the opponent plays the equilibrium strategy, a pair of quasi-Nash equilibrium only requires the first-order stationarity condition to be satisfied in their respective optimization problem. Hence quasi-Nash equilibrium can be viewed as a relaxation of Nash equilibrium.

Definition 6.2.14 (Minty Nash equilibrium (MNE))**.** *A solution pair* $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ *is said to be in Minty Nash equilibrium if and only if*

$$\langle \nabla_x \theta_x(x, y^*), x - x^* \rangle \geq 0, \forall x \in \mathcal{X}, \quad \langle \nabla_y \theta_y(x^*, y), y - y^* \rangle \geq 0, \forall y \in \mathcal{Y}.$$

The third definition given above pertains to the notion of Minty solution discussed thus far. It requires x^*/y^* to be a Minty solution of $\theta_x(\cdot, y^*)/\theta_y(x^*, \cdot)$ for fixed y^*/x^*. By the discussion in the previous section, the set of Minty solutions is only a subset of the global solutions, therefore the Minty Nash equilibrium defines a stronger concept of equilibrium than the usual notion of Nash equilibrium.

It is straightforward to conclude the following relation among these three different notions of equilibria:

$$\text{MNE} \Longrightarrow \text{NE} \Longrightarrow \text{QNE}.$$

If the objective functions possess an additional property known as *block multiconvex*, i.e. $\theta_x(\cdot, y)$ is convex for fixed $y \in \mathcal{Y}$ and $\theta_y(x, \cdot)$ is convex for fixed $x \in \mathcal{X}$, then the above relation becomes:

$$\text{MNE} = \text{NE} = \text{QNE}.$$

Let us now consider the connections among the above notions of equilibria to the solutions in the VI formulation of the two-player game (6.2.13):

$$F(z) := \begin{pmatrix} \nabla_x \theta_x(x, y) \\ \nabla_y \theta_y(x, y) \end{pmatrix}, \quad z := (x, y)^\top, \quad \mathcal{Z} := \mathcal{X} \times \mathcal{Y}, \tag{6.2.14}$$

which can be expressed as the VI model $\text{VI}(F; \mathcal{Z})$.

If $z^* = (x^*, y^*) \in \text{Sol}(\text{VI}(F; \mathcal{Z}))$, i.e.

$$\langle F(z^*), z - z^* \rangle = \langle \nabla_x \theta_x(x^*, y^*), x - x^* \rangle + \langle \nabla_y \theta_y(x^*, y^*), y - y^* \rangle \geq 0, \quad \forall z \in \mathcal{Z},$$

it is obvious that (x^*, y^*) is a pair of quasi-Nash equilibrium of the original two-player game by taking $x = x^*$ and $y = y^*$ in the above inequality. On the other hand, if (x^*, y^*) is a pair of quasi-Nash equilibrium, then the above inequality holds trivially and $z^* = (x^*, y^*) \in \mathrm{Sol}(\mathrm{VI}(F; \mathcal{Z}))$. Therefore, quasi-Nash equilibrium of the game is equivalent to the (strong) solution to the VI formulation.

Now if $z^* = (x^*, y^*) \in \mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{Z}))$, i.e.

$$\langle F(z), z - z^* \rangle = \langle \nabla_x \theta_x(x, y), x - x^* \rangle + \langle \nabla_y \theta_y(x, y), y - y^* \rangle \geq 0, \quad \forall z \in \mathcal{Z}. \tag{6.2.15}$$

By taking any arbitrary $x \in \mathcal{X}$ and $y = y^*$ in the above inequality, we have

$$\langle \nabla_x \theta_x(x, y^*), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X}.$$

Similarly we have

$$\langle \nabla_y \theta_y(x^*, y), y - y^* \rangle \geq 0, \quad \forall y \in \mathcal{Y}.$$

The above two inequalities combined indicates that (x^*, y^*) is a pair of Minty Nash equilibrium of the original two-player game. However, we note that the opposite direction is in general not true, since a Minty solution z^* to $\mathrm{VI}(F; \mathcal{X})$ requires the inequality (6.2.15) to be satisfied with $\nabla_x \theta_x(x, y)/\nabla_y \theta_y(x, y)$ while the Minty Nash equilibrium only defines on $\nabla_x \theta_x(x, y^*)/\nabla_y \theta_y(x^*, y)$, in which the opponent's strategy is fixed to be the equilibrium strategy.

We can include the two solution concepts in the VI formulation, $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{Z}))$ and $\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{Z}))$, in the previous relation and obtain

$$\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{Z})) \Longrightarrow \mathrm{MNE} \Longrightarrow \mathrm{NE} \Longrightarrow \mathrm{QNE} = \mathrm{Sol}(\mathrm{VI}(F; \mathcal{Z})).$$

In the case where both $\theta_x(x, y)$ and $\theta_y(x, y)$ are bock multiconvex, then

$$\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{Z})) \Longrightarrow \mathrm{MNE} = \mathrm{NE} = \mathrm{QNE} = \mathrm{Sol}(\mathrm{VI}(F; \mathcal{Z})).$$

Note that it is not sufficient to state the equivalence between the Minty solution in VI and others even if the objective functions are block multiconvex. However, the above relation does offer a quick argument for the existence of Nash equilibrium for non-cooperative games where the payoff functions are block multiconvex and continuously differentiable, and the constraints are convex compact sets. Indeed, the latter two conditions are exactly given in Assumption 6.2.1, which guarantees that $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{Z})) \neq \emptyset$. The functions being block multiconvex indicates that NE= $\mathrm{Sol}(\mathrm{VI}(F; \mathcal{Z}))$, which proves the existence of Nash equilibrium. This conclusion is summarized in the next proposition.

Proposition 6.2.15. *For the two-player game (6.2.13), if both $\theta_x(x, y)$ and $\theta_y(x, y)$ are block multiconvex and continuously differentiable, and $\mathcal{X}, \mathcal{Y}$ are convex compact sets, then a Nash equilibrium exists.*

Remark 6.2.16. *Indeed, the celebrated Nash Theorem has shown that a mixed strategy Nash equilibrium exists in n-player multilinear games using Brouwer's fixed-point theorem. The above discussion only points out an alternative route for showing the same conclusion with a potentially lighter algebraic derivation. The equivalence between NE and QNE as well as between QNE and* $\mathrm{Sol}(VI(F; \mathcal{Z}))$ *is straightforward. Proving* $\mathrm{Sol}(VI(F; \mathcal{Z})) \neq \emptyset$ *can be accomplished via the same Brouwer's fixed-point theorem or degree theory [19], where the details are omitted here.*

We note that restricting the number of players to be two in this section is only for the purpose of clear illustrations of the ideas. All the discussions can be easily extended to general n-player games.

6.3 Algorithm-Based Conditions on VI problems

In the discussion thus far, we have focused on non-monotone VI with the condition $\mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{X})) \neq \emptyset$, and we show that it is sufficient for a projection-type method such as ARE to converge globally with a guaranteed rate. In this section, we continue to ask the question: Are there other sufficient conditions different from the existing ones that are able to guarantee the convergence of algorithms of certain class? It turns out that through an *algorithm-based* approach, it is possible to characterize structures of the VI model by deriving sufficient conditions for which the algorithms converge. These VI problems can be of special interest since they are not necessarily monotone or satisfy the Minty condition, nonetheless the algorithms converge regardless. In particular, we present conditions on VI models based on projection-type methods, analyze their convergence behavior, and provide examples of problems satisfying these conditions.

6.3.1 Conditions for projection-type methods

In order to present the conditions to be introduced later with more precise expressions, let us first define two projection-type mappings, which play a central role in these conditions since the purpose is to characterize VI problems with guaranteed convergence for projection-type methods.

Definition 6.3.1 (Gradient projection mapping)**.** *For a given $t > 0$, assuming $\mathcal{X}$ is a closed convex set, define the "gradient projection mapping" as*

$$M(x;t) := \mathrm{Proj}_{\mathcal{X}}(x - tF(x)). \tag{6.3.16}$$

Note that the term "gradient" follows the convention in optimization, while in general F can be any vector mapping that is not necessarily a gradient mapping.

It is a well-known fact that for a fixed $t > 0$, $x^* \in \mathcal{X}$ is a fixed point of the gradient projection mapping $M(\cdot;t)$ if and only if $x^* \in \mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$. It is then natural to use a third merit function other than the (dual) gap function introduced in Section 6.2.1.2:

$$P(x) := \|M(x;t) - x\|^2.$$

We summarize the above observations in the next proposition and provide a proof for completeness.

Proposition 6.3.2. *$x^* = M(x^*;t)$ if and only if $x^* \in \mathrm{Sol}(\mathit{VI}(F;\mathcal{X}))$. Therefore, $P(x) = 0$ if and only if $x \in \mathrm{Sol}(\mathit{VI}(F;\mathcal{X}))$.*

Proof. *By the optimality condition of the projection operation, we have*

$$(y - M(x;t))^\top (M(x;t) - x + tF(x)) \geq 0, \ \ \forall y \in \mathcal{X}. \tag{6.3.17}$$

If $M(x;t) \neq x$, then by setting $y = x$ in (6.3.17) we observe

$$(M(x;t) - x)^\top F(x) \leq -\frac{1}{t}\|M(x;t) - x\|^2 < 0$$

implying that $x \notin \mathrm{Sol}(\mathit{VI}(F;\mathcal{X}))$. On the other hand, if $M(x;t) = x$, then (6.3.17) yields

$$t(y - x)^\top F(x) \geq 0, \ \ \forall y \in \mathcal{X},$$

and so $x \in \mathrm{Sol}(\mathit{VI}(F;\mathcal{X}))$. □

Remark 6.3.3. *In the literature, the mapping $M(x;1) - x$ is also referred to as the "natural map" and solving $M(x;1) - x = 0$ can be used as an equation reformulation of the VI model $\mathit{VI}(F;\mathcal{X})$. In our discussion, we do not explicitly adopt the equation reformulation approach, but only use $P(x)$ as one of the measurements of convergence.*

In view of the gradient projection mapping defined earlier, let us also define the following "extra-gradient projection mapping", which expresses the extra-gradient-type methods and the sufficient condition for convergence more succinctly.

Definition 6.3.4 (Extra-gradient projection mapping)**.** *For a given $t > 0$, assuming $\mathcal{X}$ is a closed convex set, define the "extra-gradient mapping" as*

$$M^+(x;t) := \mathrm{Proj}_{\mathcal{X}}(x - tF(M(x;t))),$$

where $M(x;t)$ is the gradient projection mapping defined in (6.3.16)*.*

We are now ready to introduce conditions that can guarantee the convergence for different projection-type methods. These conditions provide additional characterizations of the structure of a VI problem when in general we do not assume monotonicity nor Minty condition.

Definition 6.3.5. *Condition (Local Minty).*

For some fixed $t > 0$ and any $x \in \mathcal{X}$ there is $x^* \in \mathrm{Sol}(VI(F;\mathcal{X}))$ such that

$$\langle F(x), x - x^* \rangle \geq 0,$$

and for the same x^* the above inequality also holds if we replace x by $M(x;t)$.

Definition 6.3.6. *Condition (Local Minty+).*

For some fixed $t > 0$ and any $x \in \mathcal{X}$ there is $x^* \in \mathrm{Sol}(VI(F;\mathcal{X}))$ such that

$$\langle F(M(x;t)), M(x;t) - x^* \rangle \geq 0,$$

and for the same x^* the above inequality also holds if we replace x by $M^+(x;t)$.

Definition 6.3.7. *Condition (Local Minty*).*

For some fixed $t > 0$ and any $x \in \mathcal{X}$ there is $x^* \in \mathrm{Sol}(VI(F;\mathcal{X}))$ such that

$$\langle F(x), M(x;t) - x^* \rangle \geq 0,$$

and for the same x^* the above inequality also holds if we replace x by $M(x;t)$.

Definition 6.3.8. *Condition (GP).*

For some fixed $\delta > 0$ and $t > 0$, and any $x \in \mathcal{X}$ there is $x^* \in \mathrm{Sol}(VI(F;\mathcal{X}))$ such that

$$4(1+\delta)t\langle F(M(x;t)), M(x;t) - x^* \rangle + \|M(x;t) - x\|^2 \geq 0,$$

and for the same x^* the above inequality also holds if we replace x by $M(x;t)$.

Definition 6.3.9. *Condition (GP+).*

For some fixed $\delta > 0$ and $t > 0$, and any $x \in \mathcal{X}$ there is $x^* \in \mathrm{Sol}(VI(F;\mathcal{X}))$ such that

$$4(1+\delta)t\langle F(M(x;t)), M(x;t) - x^* \rangle + \|M(x;t) - x\|^2 \geq 0,$$

and for the same x^* the above inequality also holds if we replace x by $M^+(x;t)$.

Definition 6.3.10. *Condition (GP*).*

For some fixed $\delta > 0$ and $t > 0$, and any $x \in \mathcal{X}$ there is $x^* \in \mathrm{Sol}(VI(F; \mathcal{X}))$ such that

$$2(1+\delta)t\langle F(x), M(x;t) - x^*\rangle + \|M(x;t) - x\|^2 \geq 0,$$

and for the same x^* the above inequality also holds if we replace x by $M(x;t)$.

We make the following remarks on the conditions introduced above.

Remark 6.3.11.

- *Condition (Local Minty) (Local Minty*): Unlike other conditions that are seen before, both conditions are defined on sequences in $\mathcal{X}$ rather than arbitrary points. In particular these sequences are generated by the gradient projection mapping $M(x;t)$. For any specific sequence, there is a "local Minty solution" for which the inequality defined in the respective condition continues to hold.*
- *Condition (Local Minty+): When generating a sequence from the extra-gradient projection mapping $M^+(x;t)$, we immediately obtain another sequence that maps the previous sequence to their gradient projection mapping $M(x;t)$. (Local Minty+) is defined on the latter sequences.*
- *Condition (GP), (GP+), (GP*): These three conditions can be viewed as relaxations of Condition (Local Minty), (Local Minty+), and (Local Minty*), by allowing a positive term $P(x)$ in the defining inequality.*

The conditions introduced above have one property in common: they all require the defining inequality to hold for the whole sequence generated from some pre-determined mapping. In other words, a solution $x^* \in \mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$ can "attract" some $x \in \mathcal{X}$ following the particular sequence. Condition (Local Minty)/(Local Minty+)/(Local Minty*) (Definition 6.3.5, 6.3.6, and 6.3.7) guarantee the existence of such "local Minty solution" x^* with respect to arbitrary $x \in \mathcal{X}$, while Condition (GP)/(GP+)/(GP*) (Definition 6.3.8, 6.3.9, and 6.3.10) relax the previous three conditions. The term "local" is in contrast to the normal Minty solution, which is "global" since any $x^* \in \mathrm{Sol}_m(\mathrm{VI}(F; \mathcal{X}))$ is able to attract *every* point in $\mathcal{X}$ in terms of the Minty inequality $\langle F(x), x - x^*\rangle \geq 0$ in the definition. This property turns out to be critical to derive these algorithm-based conditions, which helps establish convergence of projection-type methods for those problems with more general structures than the common monotonicity or Minty condition.

Remark 6.3.12.

- *If F is monotone, then Condition (Local Minty) and Condition (Local Minty+) hold trivially (therefore so do Conditions (GP) and (GP+)).*
- *If $\mathrm{Sol}_m(VI(F;\mathcal{X})) \neq \emptyset$, then Condition (Local Minty) and Condition (Local Minty+) hold trivially (therefore so do Condition (GP) and (GP+)).*
- *It is possible that F is monotone but Conditions (Local Minty*) and (GP*) do not hold. Conversely, it is also possible that Conditions (Local Minty*) and (GP*) hold but F is not monotone or $\mathrm{Sol}_m(VI(F;\mathcal{X})) = \emptyset$.*

The next two examples demonstrate problem instances where $\mathrm{Sol}_m(\mathrm{VI}(F)) = \emptyset$ while Condition (GP)/(GP+)/(GP*) still hold.

Example 6.3.13. *Consider $\mathcal{X} = [-1, 1]$, $F(x) = -x$. In that case, $\mathrm{Sol}(VI(F;\mathcal{X})) = \{-1, 0, +1\}$ and $\mathrm{Sol}_m(VI(F;\mathcal{X})) = \emptyset$. Note that*

$$M(x;t) = \begin{cases} \min\{1, (1+t)x\}, & \text{if } x > 0 \ ; \\ 0, & \text{if } x = 0 \ ; \\ \max\{-1, (1+t)x\}, & \text{if } x < 0 \ . \end{cases}$$

In particular, for $x > 0$ we choose $x^* = 1$; for $x < 0$ we choose $x^* = -1$; for $x = 0$ we choose $x^* = 0$. It is now easy to verify that Condition (GP)/(GP+)/(GP) hold in this case. In fact, Condition (Local Minty)/(Local Minty+)/(Local Minty*) all hold in this example.*

Example 6.3.14. *Consider $\mathcal{X} = \|x\| \leq 1$, $F(x) = Qx$, where $Q = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$. In this case, $\mathrm{Sol}(VI(F;\mathcal{X})) = \left\{(1,0)^\top, (0,0)^\top, (-1,0)^\top\right\}$. None of them is a Minty solution, but we can show that $VI(F;\mathcal{X})$ satisfies (Local Minty)/(Local Minty+)/(Local Minty+) for any fixed $t \in (0, 1]$. For the first two conditions, it suffices to provide the following two observations as the proof. Also note that we can focus on $x_1 \geq 0$, since the behavior is symmetric for the case $x_1 \leq 0$.*

1. *For any $x = (x_1, x_2)^\top \in \mathcal{X}$ such that $x_1 \geq 0$, the whole sequence generated by $M(x;t)$ and $M^+(x;t)$ will remain $x_1 \geq 0$. This can be easily verified.*
2. *For any $x = (x_1, x_2)^\top \in \mathcal{X}$ such that $x_1 \geq 0$, we have*

$$\langle F(x), x - x^* \rangle \geq 0,$$

for $x^* = (1,0)^\top$. Since the above inequality results in $x_2^2 \geq x_1(x_1 - 1)$, which always holds for any $0 \leq x_1 \leq 1$.

It remains to show that Condition (Local Minty) also holds. Similarly, let us focus on* $x_1 \geq 0$ *and use* $x^* = (1,0)^\top$*. Let us denote* $x^+ = x - tF(x) = ((1+t)x_1, (1-t)x_2)^\top$*, then we have*

$$M(x;t) = \begin{cases} x^+, & \text{if } \|x^+\| \leq 1 \ ; \\ x^+/\|x^+\|, & \text{if } \|x^+\| > 1 \ . \end{cases}$$

For the case $\|x^+\| \leq 1$*, the condition*

$$\langle F(x), M(x;t) - x^* \rangle \geq 0 \tag{6.3.18}$$

reduces to

$$(1-t)x_2^2 \geq (1+t)x_1^2 - x_1,$$

where the RHS is always non-positive for $x_1 \leq \frac{1}{1+t}$*, which is the case for* $\|x^+\| \leq 1$*. Therefore inequality* (6.3.18) *holds. For* $\|x^+\| > 1$*, condition* (6.3.18) *can be reduced to*

$$(1-t)x_2^2 \geq (1+t)x_1^2 - x_1 \cdot \|x^+\|.$$

Since $\|x^+\| \geq (1+t)x_1$*, the RHS is always non-positive and condition* (6.3.18) *holds.*

The next example shows that, even if F is monotone, Condition (GP*) does not necessarily hold. Otherwise, since Condition (GP*) is sufficient for the gradient projection method to converge (as will be shown in the next section), monotonicity would have been sufficient for the convergence as well (which is not the case for the gradient projection method).

Example 6.3.15. *Consider* $\mathcal{X} = \|x\| \leq 1$*,* $F(x) = Qx$*, where* $Q = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$*. In this case, Sol(VI(*$F; \mathcal{X}$*))* $= \left\{(0,0)^\top\right\}$*. This problem is originated from the saddle point problem* $\min\max_{\|x\|^2+\|y\|^2 \leq 1} xy$ *and is monotone. For a small* $\epsilon > 0$*, consider* $x = (\epsilon, 0)^\top$*, where* $F(x) = (0, -\epsilon)^\top$*.*

Consider first $t \leq \frac{\sqrt{1-\epsilon^2}}{\epsilon}$*. In this case,* $M(x;t) = x - tF(x) = \epsilon \cdot (1,t)^\top$*. Therefore,*

$$2(1+\delta)t\langle F(x), M(x;t) - x^* \rangle + \|M(x;t) - x\|^2 = -2(1+\delta)t^2\epsilon^2 + t^2\epsilon^2 < 0.$$

On the other hand, if $t > \frac{\sqrt{1-\epsilon^2}}{\epsilon}$*, then* $M(x;t) = (1,t)^\top \cdot (1+t^2)^{-\frac{1}{2}}$*, and*

$$\|M(x;t) - x\|^2 = \left\| \frac{(1 - \epsilon\sqrt{1+t^2}, t)^\top}{\sqrt{1+t^2}} \right\|^2 = 1 + \epsilon^2 - \frac{2\epsilon}{\sqrt{1+t^2}} \leq 1 + \epsilon^2,$$

whereas

$$2(1+\delta)t\langle F(x), M(x;t) - x^* \rangle = -2(1+\delta)\epsilon \cdot t^2 \cdot (1+t^2)^{-\frac{1}{2}} < -2(1+\delta) \cdot (1-\epsilon^2),$$

where the last inequality we take $t = \frac{\sqrt{1-\epsilon^2}}{\epsilon}$*. It is then clear that for small enough* ϵ*, Condition (GP*) will not hold, even if* F *is monotone.*

6.3.2 Convergence of projection-type methods

In the previous section, we present several algorithm-based conditions that are defined for sequences generated from either the gradient projection mapping or the extra-gradient projection mapping. In this section we show how these conditions are applied in the convergence analysis for two projection-type methods, the vanilla gradient projection method and the extra-gradient method.

6.3.2.1 The gradient projection method

Consider the gradient projection method:

$$x^{k+1} \quad := \quad \arg\min_{x\in\mathcal{X}} \langle F(x^k), x - x^k\rangle + \frac{1}{2t}\|x - x^k\|^2,$$

or equivalently written as

$$x^{k+1} \quad := \quad M(x^k; t).$$

It is now clear why the conditions (Local Minty)/(Local Minty*) and their relaxations (GP)/(GP*) are defined for sequences generated from the gradient projection mapping. They assume that for each sequence generated by the gradient projection method there exists at least one solution $x^* \in \mathrm{Sol}(\mathrm{VI}(F;\mathcal{X}))$ such that their respective defining inequalities continue to hold. We first derive a key intermediate inequality from the gradient projection update itself, which makes it clearer the exact condition to be used in the following convergence analysis.

Lemma 6.3.16. *For the gradient projection method, we have*

$$\frac{1}{2}\|x^k - x^*\|^2 \geq \frac{1}{2}\|x^{k+1} - x^*\|^2 + t\langle F(x^k), x^{k+1} - x^*\rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2. \tag{6.3.19}$$

for any $x^* \in \mathrm{Sol}(\mathit{VI}(F;\mathcal{X}))$.

Proof. *Since*

$$\langle tF(x^k) + x^{k+1} - x^k, x - x^{k+1}\rangle \geq 0 \quad \forall x \in \mathcal{X},$$

with $x = x^*$, *we have*

$$\langle tF(x^k), x^* - x^{k+1}\rangle \geq \frac{1}{2}\left(\|x^{k+1} - x^k\|^2 + \|x^{k+1} - x^*\|^2 - \|x^k - x^*\|^2\right).$$

Rearranging terms gives the result. □

In view of the inequality (6.3.19), it is straightforward that Conditions (Local Minty*) and (GP*) can provide a bound on $\langle F(x^k), x^{k+1} - x^* \rangle$ for the whole sequence with respect to some $x^* \in \mathrm{Sol}(\mathrm{VI}(F; \mathcal{X}))$, thus the convergence follows. The results are summarized in the next theorem.

Theorem 6.3.17. *Under Condition (GP*), and assume F is Lipschitz continuous with constant L, the gradient projection algorithm is convergent for VI$(F; \mathcal{X})$. Moreover,*

$$\min_{1 \le k \le N} P(x^k) = O(1/N), \quad \min_{1 \le k \le N} G(x^k) = O(1/N^{\frac{1}{2}}).$$

Proof. *In view of Lemma 6.3.16 and Condition (GP*), we have:*

$$\begin{aligned}
\frac{1}{2}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) &\geq t\langle F(x^k), x^{k+1} - x^* \rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2 \\
&\overset{(GP^*)}{\geq} -\frac{1}{2(1+\delta)}\|x^{k+1} - x^k\|^2 + \frac{1}{2}\|x^{k+1} - x^k\|^2 \\
&= \frac{\delta}{2(1+\delta)}\|x^{k+1} - x^k\|^2.
\end{aligned}$$

By Condition (GP), there exists an x^* such that the above inequality holds for $k = 1, ..., N$. Therefore, summing up the inequality for $k = 1, ..., N$, we have:

$$\sum_{k=1}^{N} \|x^{k+1} - x^k\|^2 \leq \left(1 + \frac{1}{\delta}\right) \|x^1 - x^*\|^2,$$

which implies

$$\min_{1 \le k \le N} P(x^k) \leq \frac{1}{N} \sum_{k=1}^{N} \|x^{k+1} - x^k\|^2 \leq \frac{1}{N} \cdot \left(1 + \frac{1}{\delta}\right) \|x^1 - x^*\|^2 = \mathcal{O}(1/N).$$

Moreover, we can transform the measurement in $\min_{1 \le k \le N} P(x^k)$ into $\min_{1 \le k \le N} G(x^k)$. Since

$$\langle tF(x^k) + x^{k+1} - x^k, x - x^{k+1} \rangle \geq 0 \quad \forall x \in \mathcal{X},$$

we have

$$\langle F(x^k), x^{k+1} - x \rangle \leq -\frac{1}{t}(x^{k+1} - x)^\top (x^{k+1} - x^k) \leq \frac{D}{t}\|x^{k+1} - x^k\|.$$

Therefore,

$$\begin{aligned}
\langle F(x^{k+1}), x^{k+1} - x \rangle &= \langle F(x^k), x^{k+1} - x \rangle + \langle F(x^{k+1}) - F(x^k), x^{k+1} - x \rangle \\
&\leq \frac{D}{t}\|x^{k+1} - x^k\| + \|F(x^{k+1}) - F(x^k)\| \cdot \|x^{k+1} - x\| \\
&\leq D\left(\frac{1}{t} + L\right) \|x^{k+1} - x^k\|.
\end{aligned}$$

Define $k_N := \arg\min_{1\le k\le N} P(x^k)$*, then*

$$\begin{aligned}\langle F(x^{k_N+1}), x^{k_N+1} - x\rangle &\le D\left(\frac{1}{t}+L\right)\|x^{k_N+1}-x^{k_N}\| \\ &\le D\left(\frac{1}{t}+L\right)\frac{1}{N^{\frac{1}{2}}}\left(1+\frac{1}{\delta}\right)^{\frac{1}{2}}\|x^1-x^*\| \\ &\le D^2\left(\frac{1}{t}+L\right)\frac{1}{N^{\frac{1}{2}}}\left(1+\frac{1}{\delta}\right)^{\frac{1}{2}},\end{aligned}$$

which holds for all $x \in \mathcal{X}$*. This implies*

$$\min_{1\le k\le N} G(x^k) = \max_{x\in\mathcal{X}}\langle F(x^{k_N}), x^{k_N} - x\rangle \le D^2\left(\frac{1}{t}+L\right)\frac{1}{N^{\frac{1}{2}}}\left(1+\frac{1}{\delta}\right)^{\frac{1}{2}} = \mathcal{O}(1/N^{\frac{1}{2}}).$$

□

It is well-known that if F is strongly monotone, gradient projection method can be guaranteed to converge with iteration complexity $\mathcal{O}\left(\frac{L^2}{\mu^2}\log\frac{1}{\epsilon}\right)$, while F being merely monotone is insufficient for the convergence. On the other hand, Theorem 6.3.17 provides a different sufficient condition (GP*) such that gradient projection method converges globally in terms of $\min_{1\le k\le N} G(x^k) = G(x^{k_N})$. As shown in the previous examples, there exist problems which are not monotone but Condition (GP*) is satisfied.

6.3.3 The extra-gradient method

Consider

$$\begin{cases} x^{k+0.5} &:= \arg\min_{x\in\mathcal{X}}\langle F(x^k), x-x^k\rangle + \frac{1}{2t}\|x-x^k\|^2 \\ x^{k+1} &:= \arg\min_{x\in\mathcal{X}}\langle F(x^{k+0.5}), x-x^k\rangle + \frac{1}{2t}\|x-x^k\|^2, \end{cases}$$

or it can be equivalently written as

$$\begin{cases} x^{k+0.5} &:= M(x^k;t) \\ x^{k+1} &:= M^+(x^k;t), \end{cases}$$

Note that the extra-gradient method can be viewed as a special case of the ARE update discussed in Section 6.2.2 with $p = 1$. Below we introduce a key inequality derived from the update, which also appears with a similar form in the analysis for ARE (cf. (6.2.7)). Both the notation and the parameter constraint are slightly adjusted in the following lemma, therefore a proof is provided for completeness.

Lemma 6.3.18. *For the extra-gradient method, assume that F is Lipschitz continuous with constant L, and $t \le \frac{1}{\sqrt{2}L}$, the following inequality holds:*

$$\langle F(x^{k+0.5}), x^{k+0.5} - x\rangle + \frac{1}{4t}\|x^{k+0.5} - x^k\|^2 \le \frac{1}{2t}\left[\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right]. \tag{6.3.20}$$

Proof. *Optimality of $x^{k+0.5}$:*

$$\langle F(x^k) + \frac{1}{t}(x^{k+0.5} - x^k), x - x^{k+0.5}\rangle \ge 0, \ \forall x \in \mathcal{X}. \tag{6.3.21}$$

Substituting $x = x^{k+1}$ in (6.3.21) we have

$$\begin{aligned} & \langle F(x^k), x^{k+1} - x^{k+0.5}\rangle \\ \ge\ & \frac{1}{t}\langle x^{k+0.5} - x^k, x^{k+0.5} - x^{k+1}\rangle \\ =\ & \frac{1}{2t}\left(\|x^{k+0.5} - x^k\|^2 + \|x^{k+1} - x^{k+0.5}\|^2 - \|x^{k+1} - x^k\|^2\right). \end{aligned} \tag{6.3.22}$$

On the other hand, by the optimality condition at x^{k+1} we have

$$\langle F(x^{k+0.5}) + \frac{1}{t}(x^{k+1} - x^k), x - x^{k+1}\rangle \ge 0, \ \textit{for all } x \in \mathcal{X}.$$

Hence,

$$\begin{aligned} & \langle F(x^{k+0.5}), x - x^{k+1}\rangle \\ \ge\ & \frac{1}{t}\langle x^{k+1} - x^k, x^{k+1} - x\rangle \\ =\ & \frac{1}{2t}\left(\|x^{k+1} - x\|^2 + \|x^{k+1} - x^k\|^2 - \|x^k - x\|^2\right), \ \textit{for all } x \in \mathcal{X}. \end{aligned} \tag{6.3.23}$$

Continue with the above inequality, for any given $x \in \mathcal{X}$ we have

$$\begin{aligned} & \frac{1}{2t}\left(\|x^{k+1} - x\|^2 + \|x^{k+1} - x^k\|^2 - \|x^k - x\|^2\right) \\ \overset{(6.3.23)}{\le}\ & \langle F(x^{k+0.5}), x - x^{k+1}\rangle \\ =\ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \langle F(x^{k+0.5}), x^{k+0.5} - x^{k+1}\rangle \\ =\ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \langle F(x^{k+0.5}) - F(x^k), x^{k+0.5} - x^{k+1}\rangle + \langle F(x^k), x^{k+0.5} - x^{k+1}\rangle \\ \le\ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \|F(x^{k+0.5}) - F(x^k)\| \cdot \|x^{k+0.5} - x^{k+1}\| \\ & + \langle F(x^k), x^{k+0.5} - x^{k+1}\rangle \\ \le\ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \frac{t\|F(x^{k+0.5}) - F(x^k)\|^2}{2} + \frac{\|x^{k+0.5} - x^{k+1}\|^2}{2t} \\ & + \langle F(x^k), x^{k+0.5} - x^{k+1}\rangle \\ \le\ & \langle F(x^{k+0.5}), x - x^{k+0.5}\rangle + \frac{tL^2\|x^{k+0.5} - x^k\|^2}{2} + \frac{\|x^{k+0.5} - x^{k+1}\|^2}{2t} \\ & + \langle F(x^k), x^{k+0.5} - x^{k+1}\rangle. \end{aligned}$$

Since $t \leq \frac{1}{\sqrt{2}L}$, *and with* (6.3.22) *we have*

$$\begin{aligned}
& \frac{1}{2t}\left(\|x^{k+1}-x\|^2+\|x^{k+1}-x^k\|^2-\|x^k-x\|^2\right) \\
\leq\ & \langle F(x^{k+0.5}), x-x^{k+0.5}\rangle + \frac{\|x^{k+0.5}-x^k\|^2}{4t} + \frac{\|x^{k+0.5}-x^{k+1}\|^2}{2t} \\
& +\frac{1}{2t}\left[-\|x^{k+0.5}-x^k\|^2-\|x^{k+1}-x^{k+0.5}\|^2+\|x^{k+1}-x^k\|^2\right].
\end{aligned}$$

Canceling out terms, we simplify the above inequality into

$$\langle F(x^{k+0.5}), x^{k+0.5}-x\rangle + \frac{1}{4t}\|x^{k+0.5}-x^k\|^2 \leq \frac{1}{2t}\left[\|x^k-x\|^2-\|x^{k+1}-x\|^2\right].$$

□

One can immediately identify the connections between inequality (6.3.20) and Condition (GP+). In fact, inequality (6.3.20) plays the central role in the convergence of the extra-gradient method (or in general, the extra-gradient-type method such as ARE), and iteration complexities of different orders can be established following this inequality based on the conditions imposed on the VI model. The most conventional assumption will be the (strong) monotonicity of F, while in Section 6.2.2 it is relaxed to be Minty condition $\mathrm{Sol}_m(\mathrm{VI}(F;\mathcal{X})) \neq \emptyset$. Here, Condition (GP+) provides a more direct way to guide the convergence analysis of the extra-gradient method based on inequality (6.3.20), as summarized in the next theorem.

Theorem 6.3.19. *Under Condition (GP+), and assume F is Lipschitz continuous with constant L, the extra-gradient method with $t \leq \frac{1}{\sqrt{2}L}$ is convergent for VI$(F;\mathcal{X})$. Moreover,*

$$\min_{1\leq k\leq N} P(x^k) = O(1/N), \qquad \min_{1\leq k\leq N} G(x^{k+0.5}) = O(1/N^{\frac{1}{2}}).$$

Proof. *In view of Lemma 6.3.18 and Condition (GP+), we have:*

$$\begin{aligned}
\frac{1}{2t}\left[\|x^k-x^*\|^2-\|x^{k+1}-x^*\|^2\right] &\geq \langle F(x^{k+0.5}), x^{k+0.5}-x^*\rangle + \frac{1}{4t}\|x^{k+0.5}-x^k\|^2 \\
&\overset{(GP+)}{\geq} -\frac{1}{4t(1+\delta)}\|x^{k+0.5}-x^k\|^2 + \frac{1}{4t}\|x^{k+0.5}-x^k\|^2 \\
&= \frac{\delta}{1+\delta}\frac{1}{4t}\|x^{k+0.5}-x^k\|^2.
\end{aligned}$$

Since Condition (GP+) also asserts that there exists an x^* such that the above inequality holds for $k=1,...,N$, summing up the inequality for $k=1,...,N$ gives us:

$$\sum_{k=1}^{N}\|x^{k+0.5}-x^k\|^2 \leq 2\left(1+\frac{1}{\delta}\right)\|x^1-x^*\|^2,$$

which implies

$$\min_{1\le k\le N} P(x^k) \le \frac{1}{N}\sum_{k=1}^{N}\|x^{k+0.5}-x^k\|^2 \le \frac{2}{N}\cdot\left(1+\frac{1}{\delta}\right)\|x^1-x^*\|^2 = \mathcal{O}(1/N).$$

Moreover, we can transform the measurement in $\min_{1\le k\le N} P(x^k)$ *into* $\min_{1\le k\le N} G(x^{k+0.5})$*. Since*

$$\langle tF(x^k)+x^{k+0.5}-x^k, x-x^{k+0.5}\rangle \ge 0 \quad \forall x\in\mathcal{X},$$

we have

$$\langle F(x^k), x^{k+0.5}-x\rangle \le -\frac{1}{t}(x^{k+0.5}-x)^\top(x^{k+0.5}-x^k) \le \frac{D}{t}\|x^{k+0.5}-x^k\|.$$

Therefore,

$$\begin{aligned}\langle F(x^{k+0.5}), x^{k+0.5}-x\rangle &= \langle F(x^k), x^{k+0.5}-x\rangle + \langle F(x^{k+0.5})-F(x^k), x^{k+0.5}-x\rangle \\ &\le \frac{D}{t}\|x^{k+0.5}-x^k\| + \|F(x^{k+0.5})-F(x^k)\|\cdot\|x^{k+0.5}-x\| \\ &\le D\left(\frac{1}{t}+L\right)\|x^{k+0.5}-x^k\|.\end{aligned}$$

Define $k_N := \arg\min_{1\le k\le N} P(x^k)$*, then*

$$\begin{aligned}\langle F(x^{k_N+0.5}), x^{k_N+0.5}-x\rangle &\le D\left(\frac{1}{t}+L\right)\|x^{k_N+0.5}-x^{k_N}\| \\ &\le D\left(\frac{1}{t}+L\right)\frac{\sqrt{2}}{N^{\frac{1}{2}}}\left(1+\frac{1}{\delta}\right)^{\frac{1}{2}}\|x^1-x^*\| \\ &\le D^2\left(\frac{1}{t}+L\right)\frac{\sqrt{2}}{N^{\frac{1}{2}}}\left(1+\frac{1}{\delta}\right)^{\frac{1}{2}},\end{aligned}$$

which holds for all $x\in\mathcal{X}$*. This implies*

$$\begin{aligned}&\min_{1\le k\le N} G(x^{k+0.5}) = \max_{x\in\mathcal{X}}\langle F(x^{k_N+0.5}), x^{k_N+0.5}-x\rangle \\ \le\ & D^2\left(\frac{1}{t}+L\right)\frac{\sqrt{2}}{N^{\frac{1}{2}}}\left(1+\frac{1}{\delta}\right)^{\frac{1}{2}} = \mathcal{O}(1/N^{\frac{1}{2}}).\end{aligned}$$

□

The convergence rate in Theorem 6.3.19 turns out to be the same as the rate in Theorem 6.3.17 for gradient projection method when Condition (GP*) is satisfied instead, as well as the rate in Theorem 6.2.8 for ARE ($p = 1$) when the Minty condition is satisfied. As discussed in the earlier examples, there exists problems where Conditions (GP+) or (GP*) are satisfied but no Minty solution exists. On the other hand, while these assumptions are able to provide alternative sufficient conditions for the convergence of certain class of algorithms, in general they can be difficult to verify *a priori* due to the requirement to hold for the whole sequence.

6.4 Conclusion

In this chapter, we discuss sufficient conditions for projection-type methods to converge in VI problems that are not necessarily monotone. We first focus on the problem where a Minty solution exists, which is a relaxation of the monotonicity assumption. We derive the guaranteed global convergence rate for a general extra-gradient type method ARE under the Minty condition, and then we extend the discussion to properties and implications of Minty solutions in more specific problem classes such as optimization and Nash games. Finally, we present conditions on VI problems that are algorithm-based, in the sense that they are closely connected to the algorithms we are interested in (in particular, projection-type methods) and can suitably serve as sufficient conditions to guarantee the convergence. Conventionally, the algorithms are designed for problems where assumptions on the structure are made *a priori*, and the convergence is only guaranteed under these assumptions. In this paper, we provide an alternative aspect, by "desinging" conditions on the VI model such that they are sufficient to guarantee convergence of certain class of algorithms. We show that this approach is indeed capable of characterizing different classes of VI problems (potentially broader) from the existing ones such as monotone VI or VI with Minty solutions. We analyze the convergence of gradient projection method and extra-gradient method under the proposed conditions. There are still questions remaining, such as: if there are other algorithm-based conditions that can be derived for different projection-type methods or non-projection-type methods; if there exist different characterizations of these conditions such that they can be more easily verified. Answering these questions require some efforts in the future research.

Chapter 7

Conclusions and Discussions

In this dissertation, we studied variational inequality problems from the perspectives of designing optimal algorithms and analyzing their iteration complexities. Many of these algorithms are not just designed to retain optimal theoretical guarantee of iteration complexities, but they aim to provide additional insights into underlying connections among other existing optimal algorithms in the literature. For the first-order algorithms, we developed a general principle, namely the extra-point approach, to systematically guide the generation of optimal methods that match the lower bound. These methods could include many existing optimal methods in the literature, and they may also be extended to other unseen methods given more specific problem structures at hand. For the high-order algorithms, we discovered that the key to designing optimal algorithms is to ensure a (p^{th}-order) Lipschitz bound is satisfied between the two operators used in subsequent extra-gradient (type) updates in each iteration. Therefore, any "approximation" operator satisfying such bound, coupled with an appropriate regularization term, can be used in an high-order optimal algorithm. A unifying framework is thus proposed based on such concept of approximation, which not only covers many existing optimal high-order methods but also opens up the possibilities for new algorithms that target the specific problem structures.

We also studied the stochastic algorithms for VI, where the operator in a problem can only be accessed via noisy estimations due to various practical reasons. We showed that the stochastic extensions of two new algorithms, generated from the extra-point approach, can achieve the optimal iteration complexity when the stochastic errors are simultaneously reduced throughout the iterations, and we further discussed the application of these algorithms to black-box saddle point problems as stochastic zeroth-order approaches, together with a sample iteration complexity analysis. In addition to the stochasticity due to estimation of the operator itself, a different layer of stochasticity can result from the finite-sum

structure in a VI problem, which has received increasing attention in today's large-scale machine learning problems. In particular, we focused on a class of finite-sum VI problems where the (finite-sum) optimization structure is also present. Explicitly exploiting such special structure enables us to develop variance reduced algorithms with better gradient complexities compared to those designed only for general finite-sum VI problems, in both monotone or strongly monotone cases. This class of VI problems has its own interests, since the saddle-point reformulation of a (finite-sum) constrained finite-sum optimization is an immediate example problem in this class, which is commonly seen in Neyman-Pearson classification problems.

We further extended our antenna to a relatively unexplored area in VI: to design and analyze solution methods for non-monotone VI. Indeed, many interesting real-world problems are non-monotone, while common methods are proposed and guaranteed convergence only for monotone VI. Without the global structure of monotonicity, the existence of Minty solution turns out to be crucial, which characterizes a subset of (global) solutions that is able to attract each point in the constraint set following iterations of properly designed accelerated algorithms such as the extra-gradient (type) methods. Further questions arise as for whether or not there are different sufficient conditions to guarantee convergence of common projection-type algorithms other than the Minty condition. We provided insights and analysis for this question through designing algorithm-based conditions to characterize various VI problem classes. Through this approach, we no longer stick to a specific VI problem class and design algorithms within such class, but we aim to discover additional structures for VI problems where the convergence of projection-type methods can be ensured without monotonicity or even Minty condition.

As a broader problem class than optimization, VI can still have rich but rather implicit structures compared to optimization, and more in-depth investigations into these structures can indeed lead to new developments in the contemporary research in VI. While the many recent developments in optimization such as high-order methods, lower iteration complexity bounds, variance-reduced algorithms, (constrained) non-convex optimization, have all contribute to revealing similar insights and underlying structures in VI that are unexplored in the past, such revelation is by no means complete and in fact continues to introduce more interesting questions within the field. For example, for high-order methods in VI, while optimal iteration complexities can be established, there is still lack of explicit and systematic discussions for approaches to solve the subproblem efficiently, making the algorithms more conceptual than practical. VI problems with general composite structures (in the form $F = H \cdot G$) haven't receive much research attention before. While the general framework ARE is able to provide insights into designing algorithms for VI with such special structures, it can be a potential direction of research for more detailed investigation into

the structures and solution methods. Non-monotone VI, especially with no Minty solutions, marks a difficult area in the research of VI. Deriving sufficient conditions that are easier to verify to guarantee convergence for common methods, as well as exploring structures in the constraints that could help ease such difficulty may become essential for future research. Deriving duality theories and certain primal-dual methods for VI can be another interesting direction to explore in the future.

Chapter 8

Bibliography

[1] Adil, D., Bullins, B., Jambulapati, A., and Sachdeva, S. Optimal methods for higher-order smooth monotone variational inequalities. *arXiv preprint arXiv:2205.06167* (2022).

[2] Alacaoglu, A., and Malitsky, Y. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory* (2022), PMLR, pp. 778–816.

[3] Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research 18*, 1 (2017), 8194–8244.

[4] Aravkin, A. Y., Burke, J. V., Drusvyatskiy, D., Friedlander, M. P., and Roy, S. Level-set methods for convex optimization. *Mathematical Programming 174*, 1 (2019), 359–390.

[5] Bauschke, H. H., and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*, vol. 408. Springer, 2011.

[6] Beck, A., and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences 2*, 1 (2009), 183–202.

[7] Bubeck, S., Lee, Y. T., and Singh, M. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187* (2015).

[8] Bullins, B., and Lai, K. A. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. *SIAM Journal on Optimization 32*, 3 (2022), 2208–2229.

[9] Burachik, R. S., and Millan, R. D. A projection algorithm for non-monotone variational inequalities. *Set-Valued and Variational Analysis 28*, 1 (2020), 149–166.

[10] Chen, Y., Lan, G., and Ouyang, Y. Accelerated schemes for a class of variational inequalities. *Mathematical Programming 165*, 1 (2017), 113–149.

[11] Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with optimism. *arXiv preprint arXiv:1711.00141* (2017).

[12] d'Aspremont, A., Scieur, D., and Taylor, A. Acceleration methods. *arXiv preprint arXiv: 2101.09545* (2021).

[13] Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems 27* (2014).

[14] Doikov, N., and Nesterov, Y. Local convergence of tensor methods. *arXiv preprint arXiv:1912.02516* (2019).

[15] Drori, Y., and Taylor, A. On the oracle complexity of smooth strongly convex minimization. *arXiv preprint arXiv:2101.09740* (2021).

[16] Drusvyatskiy, D., Fazel, M., and Roy, S. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization 28*, 1 (2018), 251–271.

[17] Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory 61*, 5 (2015), 2788–2806.

[18] Facchinei, F., Fischer, A., and Kanzow, C. Regularity properties of a semismooth reformulation of variational inequalities. *SIAM Journal on Optimization 8*, 3 (1998), 850–869.

[19] Facchinei, F., and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems.* Springer Science & Business Media, 2007.

[20] Gao, X. *Low-order optimization algorithms: iteration complexity and applications.* Ph.D. Thesis, University of Minnesota, 2018.

[21] Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551* (2018).

[22] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems* (2014), pp. 2672–2680.

[23] Harker, P. T., and Pang, J.-S. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming 48*, 1 (1990), 161–220.

[24] He, Y. Solvability of the Minty variational inequality. *Journal of Optimization Theory and Applications 174*, 3 (2017), 686–692.

[25] Huang, K., Zhang, J., and Zhang, S. Cubic regularized Newton method for the saddle point models: A global and local convergence analysis. *Journal of Scientific Computing 91*, 2 (2022), 1–31.

[26] Huang, K., and Zhang, S. New first-order algorithms for stochastic variational inequalities. *arXiv preprint arXiv:2107.08341* (2021).

[27] Huang, K., and Zhang, S. A unifying framework of accelerated first-order approach to strongly monotone variational inequalities. *arXiv preprint arXiv:2103.15270* (2021).

[28] Huang, K., and Zhang, S. An approximation-based regularized extra-gradient method for monotone variational inequalities. *arXiv preprint arXiv:2210.04440* (2022).

[29] Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization 27*, 2 (2017), 686–724.

[30] Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization 29*, 1 (2019), 175–206.

[31] Jalilzadeh, A., and Shanbhag, U. V. A proximal-point algorithm with variable sample-sizes (PPAWSS) for monotone stochastic variational inequality problems. In *2019 Winter Simulation Conference (WSC)* (2019), IEEE, pp. 3551–3562.

[32] Jiang, H., and Xu, H. Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control 53*, 6 (2008), 1462–1475.

[33] Jiang, R., and Mokhtari, A. Generalized optimistic methods for convex-concave saddle point problems. *arXiv preprint arXiv:2202.09674* (2022).

[34] Johnson, R., and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems 26* (2013).

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

[36] Kannan, A., and Shanbhag, U. V. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications 74*, 3 (2019), 779–820.

[37] Karimi, S., and Vavasis, S. A unified convergence bound for conjugate gradient and accelerated gradient. *arXiv preprint arXiv:1605.00320* (2016).

[38] Karimi, S., and Vavasis, S. A single potential governing convergence of conjugate gradient, accelerated gradient and geometric descent. *arXiv preprint arXiv:1712.09498* (2017).

[39] Kim, D. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming 190*, 1 (2021), 57–87.

[40] Kim, D., and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical Programming 159*, 1 (2016), 81–107.

[41] Korpelevich, G. The extragradient method for finding saddle points and other problems. *Matecon 12* (1976), 747–756.

[42] Koshal, J., Nedic, A., and Shanbhag, U. V. Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control 58*, 3 (2012), 594–609.

[43] Kotsalis, G., Lan, G., and Li, T. Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation. *arXiv preprint arXiv:2011.02987* (2020).

[44] Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135* (2021).

[45] Lan, G., and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical programming 171*, 1 (2018), 167–215.

[46] Larson, J., Menickelly, M., and Wild, S. M. Derivative-free optimization methods. *arXiv preprint arXiv:1904.11585* (2019).

[47] Lee, S., and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems 34* (2021), 22588–22600.

[48] Lei, M., and He, Y. An extragradient method for solving variational inequalities without monotonicity. *Journal of Optimization Theory and Applications 188* (2021), 432–446.

[49] Liang, T., and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics* (2019), PMLR, pp. 907–915.

[50] Lieder, F. On the convergence rate of the Halpern-iteration. *Optimization letters 15*, 2 (2021), 405–418.

[51] Lin, Q., Ma, R., and Yang, T. Level-set methods for finite-sum constrained convex optimization. In *International conference on machine learning* (2018), PMLR, pp. 3112–3121.

[52] Lin, Q., Nadarajah, S., and Soheili, N. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization 28*, 4 (2018), 3290–3311.

[53] Lin, T., Jordan, M., et al. Perseus: A simple high-order regularization method for variational inequalities. *arXiv preprint arXiv:2205.03202* (2022).

[54] Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O'Reilly, U.-M. Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International conference on machine learning* (2020), PMLR, pp. 6282–6293.

[55] Martinet, B. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge 4*, R3 (1970), 154–158.

[56] Menickelly, M., and Wild, S. M. Derivative-free robust optimization by outer approximations. *Mathematical Programming 179*, 1 (2020), 157–193.

[57] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629* (2018).

[58] Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511* (2019).

[59] Mokhtari, A., Ozdaglar, A., and Pattathil, S. Convergence rate of $O(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization 30*, 4 (2020), 3230–3251.

[60] Monteiro, R. D. C., and Svaiter, B. F. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization 20*, 6 (2010), 2755–2787.

[61] Monteiro, R. D. C., and Svaiter, B. F. Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems. *SIAM Journal on Optimization 22*, 3 (2012), 914–935.

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

[63] Nemirovski, A., and Yudin, D. B. Problem complexity and method efficiency in optimization.

[64] Nemirovsky, A. S. Information-based complexity of linear operator equations. *Journal of Complexity 8*, 2 (1992), 153–175.

[65] Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady SSSR* (1983), vol. 269, pp. 543–547.

[66] Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2003.

[67] Nesterov, Y. Cubic regularization of Newton's method for convex problems with constraints.

[68] Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming 109*, 2-3 (2007), 319–344.

[69] Nesterov, Y. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming* (2018), 1–27.

[70] Nesterov, Y., and Polyak, B. T. Cubic regularization of Newton method and its global performance. *Mathematical Programming 108*, 1 (2006), 177–205.

[71] Nesterov, Y., and Scrimali, L. Solving strongly monotone variational and quasi-variational inequalities. *Available at SSRN 970903* (2006).

[72] Nesterov, Y., and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics 17*, 2 (2017), 527–566.

[73] Ostroukhov, P., Kamalov, R., Dvurechensky, P., and Gasnikov, A. Tensor methods for strongly convex strongly concave saddle point problems and strongly monotone variational inequalities. *arXiv preprint arXiv:2012.15595* (2020).

[74] Palaniappan, B., and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems* (2016), pp. 1416–1424.

[75] Pang, J.-S., and Scutari, G. Nonconvex games with side constraints. *SIAM Journal on Optimization 21*, 4 (2011), 1491–1522.

[76] Park, J., and Ryu, E. K. Exact optimal accelerated complexity for fixed-point iterations. *arXiv preprint arXiv:2201.11413* (2022).

[77] Peng, J.-M., and Fukushima, M. A hybrid Newton method for solving the variational inequality problem via the d-gap function. *Mathematical Programming 86* (1999), 367–386.

[78] Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics 4*, 5 (1964), 1–17.

[79] Popov, L. D. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR 28*, 5 (1980), 845–848.

[80] Qi, L., and Jiang, H. Semismooth Karush-Kuhn-Tucker equations and convergence analysis of Newton and quasi-Newton methods for solving these equations. *Mathematics of Operations Research 22*, 2 (1997), 301–325.

[81] Robbins, H., and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics* (1951), 400–407.

[82] Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization 14*, 5 (1976), 877–898.

[83] Rockafellar, R. T., and Sun, J. Solving monotone stochastic variational inequalities and complementarity problems by progressive hedging. *Mathematical Programming 174*, 1 (2019), 453–471.

[84] Rockafellar, R. T., and Wets, R. J.-B. Stochastic variational inequalities: single-stage to multistage. *Mathematical Programming 165*, 1 (2017), 331–360.

[85] Roy, A., Chen, Y., Balasubramanian, K., and Mohapatra, P. Online and bandit algorithms for nonstationary stochastic saddle-point optimization. *arXiv preprint arXiv:1912.01698* (2019).

[86] Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming 162*, 1 (2017), 83–112.

[87] Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and trends in Machine Learning 4*, 2 (2011), 107–194.

[88] Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research 18*, 1 (2017), 1703–1713.

[89] Shanbhag, U. V. Stochastic variational inequality problems: Applications, analysis, and algorithms. In *Theory Driven by Influential Applications*. INFORMS, 2013, pp. 71–107.

[90] Sibony, M. Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone. *Calcolo 7*, 1 (1970), 65–183.

[91] Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems 33* (2020), 14303–14314.

[92] Taji, K., Fukushima, M., and Ibaraki, T. A globally convergent Newton method for solving strongly monotone variational inequalities. *Mathematical Programming 58*, 1-3 (1993), 369–383.

[93] Taylor, A., and Drori, Y. An optimal gradient method for smooth (possibly strongly) convex minimization. *arXiv preprint arXiv:2101.09741* (2021).

[94] Tong, X., Feng, Y., and Zhao, A. A survey on Neyman-Pearson classification and suggestions for future research. *Wiley Interdisciplinary Reviews: Computational Statistics 8*, 2 (2016), 64–81.

[95] Tran-Dinh, Q., and Luo, Y. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150* (2021).

[96] Tseng, P. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics 60*, 1-2 (1995), 237–252.

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

[98] Van Scoy, B., Freeman, R. A., and Lynch, K. M. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters 2*, 1 (2017), 49–54.

[99] Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819* (2020).

[100] Xie, G., Luo, L., Lian, Y., and Zhang, Z. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *International Conference on Machine Learning* (2020), PMLR, pp. 10504–10513.

[101] Xu, T., Wang, Z., Liang, Y., and Poor, H. V. Gradient free minimax optimization: Variance reduction and faster convergence. *arXiv preprint arXiv:2006.09361* (2020).

[102] Ye, M. An infeasible projection type algorithm for nonmonotone variational inequalities. *Numerical Algorithms 89*, 4 (2022), 1723–1742.

[103] Yoon, T., and Ryu, E. K. Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning* (2021), PMLR, pp. 12098–12109.

[104] Yoon, T., and Ryu, E. K. Accelerated minimax algorithms flock together. *arXiv preprint arXiv:2205.11093* (2022).

[105] Yousefian, F., Nedić, A., and Shanbhag, U. V. A regularized smoothing stochastic approximation (RSSA) algorithm for stochastic variational inequality problems. In *2013 Winter Simulations Conference (WSC)* (2013), IEEE, pp. 933–944.

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

[107] Yousefian, F., Nedić, A., and Shanbhag, U. V. On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming 165*, 1 (2017), 391–431.

[108] Zhang, J., Hong, M., and Zhang, S. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming* (2021), 1–35.

[109] Zhou, K., Ding, Q., Shang, F., Cheng, J., Li, D., and Luo, Z.-Q. Direct acceleration of SAGA using sampled negative momentum. In *The 22nd International Conference on Artificial Intelligence and Statistics* (2019), PMLR, pp. 1602–1610.