Journal of Machine Learning Research 22 (2021) 1-34 Submitted 5/20; Revised 3/21; Published 5/21

First-order Convergence Theory for Weakly-Convex-Weakly-Concave Min-max Problems

Mingrui Liu MINGRUILIU.ML@GMAIL.COM
Department of Computer Science
The University of Iowa
Iowa City, IA, 52242, USA

Hassan Rafique HASSAN-RAFIQUE@UIOWA.EDU
Department of Mathematics
The University of Iowa
Iowa City, IA, 52242, USA

Qihang Lin QIHANG-LIN@UIOWA.EDU
Business Analytics Department
The University of Iowa
Iowa City, IA, 52242, USA

Tianbao Yang TIANBAO-YANG@UIOWA.EDU
Department of Computer Science
The University of Iowa
Iowa City, IA, 52242, USA

Editor: Zhihua Zhang

Abstract

In this paper, we consider first-order convergence theory and algorithms for solving a class of non-convex non-concave min-max saddle-point problems, whose objective function is weakly convex in the variables of minimization and weakly concave in the variables of maximization. It has many important applications in machine learning including training Generative Adversarial Nets (GANs). We propose an algorithmic framework motivated by the inexact proximal point method, where the weakly monotone variational inequality (VI) corresponding to the original min-max problem is solved through approximately solving a sequence of strongly monotone VIs constructed by adding a strongly monotone mapping to the original gradient mapping. We prove first-order convergence to a nearly stationary solution of the original min-max problem of the generic algorithmic framework and establish different rates by employing different algorithms for solving each strongly monotone VI. Experiments verify the convergence theory and also demonstrate the effectiveness of the proposed methods on training GANs.

Keywords: Weakly-Convex-Weakly-Concave, Min-max, Generative Adversarial Nets, Variational Inequality, First-order Convergence

1. Introduction

This paper is motivated by solving the following min-max saddle-point problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}), \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}^q$ are closed convex sets, and $f(\mathbf{x}, \mathbf{y})$ is real-valued, continuous, not necessarily convex in $\mathbf{x}$, and not necessarily concave in $\mathbf{y}$. This problem has broad applications in machine learning and statistics. One such example that recently attracts tremendous attention in machine learning is training Generative Adversarial Networks (GAN) where $\mathbf{x}$ denotes the parameter of the generator network and $\mathbf{y}$ denotes the parameter of the discriminator network (Goodfellow et al., 2014).

Problem (1) has been well studied when $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$ and many algorithms have been developed with provable non-asymptotic convergence properties (Nemirovski and Yudin, 1978; Nemirovski, 2005; Nemirovski et al., 2009; Chambolle and Pock, 2011; Yang et al., 2015). However, when $f(\mathbf{x}, \mathbf{y})$ is non-convex in $\mathbf{x}$ and non-concave in $\mathbf{y}$, the developments of algorithms for (1) with provable non-asymptotic convergence guarantees remain rare and most existing studies under this setting only focus on asymptotic convergence analysis (Cherukuri et al., 2017; Daskalakis et al., 2017; Heusel et al., 2017; Nagarajan and Kolter, 2017; Solodov and Svaiter, 1999; Wang et al., 2001; Bao and Khanh, 2005; Burachik and Millan, 2016). Besides non-asymptotic convergence, an important property of an algorithm for (1) is what type of solutions it can guarantee. In the studies of convex-concave saddle-point problems, one is interested in finding a *saddle point* $(\mathbf{x}_*, \mathbf{y}_*)$ that satisfies $f(\mathbf{x}_*, \mathbf{y}) \leq f(\mathbf{x}_*, \mathbf{y}_*) \leq f(\mathbf{x}, \mathbf{y}_*), \forall \mathbf{x} \in \mathcal{X}, \forall \mathbf{y} \in \mathcal{Y}$. However, finding such a saddle point for (1) is in general NP-hard (in fact, minimizing a generic non-convex function is already NP-hard (see e.g., (Hillar and Lim, 2013)). Following the existing studies on non-convex optimization, we may instead consider finding a first-order *stationary* solution to (1), which is a solution $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_*$ with

$$\mathcal{F}_* = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \,\middle|\, \begin{array}{l} 0 \in \partial_x [f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{X}}(\mathbf{x})], \\ 0 \in \partial_y [-f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Y}}(\mathbf{y})] \end{array} \right\}, \tag{2}$$

where ∂_x and ∂_y are partial subdifferentials defined in Section 3, and $1_{\mathcal{X}}(\mathbf{x})$ and $1_{\mathcal{Y}}(\mathbf{y})$ are the indicator functions of the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. It is notable that $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_*$ is the first-order necessary condition for $(\mathbf{x}, \mathbf{y})$ to be a (local) saddle point of (1). However, iterative algorithms typically do not guarantee an exact stationary solution within finitely many iterations. Hence, to establish non-asymptotic convergence of algorithms for (1), we focus on finding a *nearly ϵ-stationary* solution which is a useful notion of stationarity when a problem is non-smooth or constrained.

Although $\mathcal{F}_*$ is a direct extension of the set of first-order stationary points of minimization problems to min-max problems, it remains unclear how to prove an algorithm converges non-asymptotically to $\mathcal{F}_*$ when the min-max problem is non-convex non-concave. To the best of our knowledge, **this is the first work that proves the non-asymptotic convergence** of first-order methods to a nearly stationary solution of a class of non-smooth non-convex non-concave min-max problems. The key novelty of our analysis is viewing min-max problems through the lens of variational inequalities.

Let $F(\mathbf{z}) : \mathcal{Z} \rightrightarrows \mathbb{R}^d$ be a set-valued mapping and $\mathcal{Z} \subset \mathbb{R}^d$ be a closed convex set. The *variational inequality* (VI) problem, also known as the *Stampacchia variational inequality* (SVI) problem (Hartman and Stampacchia, 1966), associated with F and $\mathcal{Z}$, is denoted by $\mathrm{SVI}(F, \mathcal{Z})$ and concerns finding $\mathbf{z}^* \in \mathcal{Z}$ such that

$$\exists \boldsymbol{\xi}^* \in F(\mathbf{z}^*) \text{ s.t. } \langle \boldsymbol{\xi}^*, \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \ \forall \mathbf{z} \in \mathcal{Z}. \tag{3}$$

Table 1: Summary of complexity results for finding a nearly ϵ-stationary solution of the ρ-weakly-convex-weakly-concave min-max saddle-point problems (where $\rho > 0$) by using different algorithms to solve the strongly monotone subproblems. In the stochastic and deterministic settings, the complexity refers to the iteration complexity and, in the finite-sum setting, it refers to gradient complexity (i.e., the number of gradients computed). L refers to the Lipschitz constant. Note that the per-iteration cost in deterministic setting is the same as evaluating the (sub)gradient, while in the stochastic and finite-sum setting it is the same as evaluating the stochastic gradient based on one individual component function.

Setting	Algorithms for sub-problems	Lipschitz	Complexity
Stochastic	Stochastic Subgradient Method	No	$O\left(\max(\rho^6, \rho^3)/\epsilon^6\right)$
Deterministic	Subgradient Method	No	$O\left(\max(\rho^6, \rho^3)/\epsilon^6\right)$
	Gradient Descent Method	Yes	$\widetilde{O}\left(L^2/\epsilon^2\right)$
	Extragradient Method	Yes	$\widetilde{O}\left(L\rho/\epsilon^2\right)$
	Nesterov's Accelerated Method	Yes	$\widetilde{O}\left(L\rho/\epsilon^2\right)$
Finite-sum (n summands)	Variance Reduction	Yes	$\widetilde{O}\left(n\rho^2/\epsilon^2 + L^2/\epsilon^2\right)$

A closely related but different problem is the *Minty variational inequality* (MVI) problem (Minty et al., 1962) associated with F and $\mathcal{Z}$, which is denoted by MVI$(F, \mathcal{Z})$ and concerns finding $\mathbf{z}_* \in \mathcal{Z}$ such that

$$\langle \boldsymbol{\xi}, \mathbf{z} - \mathbf{z}_* \rangle \geq 0, \ \forall \mathbf{z} \in \mathcal{Z}, \forall \boldsymbol{\xi} \in F(\mathbf{z}). \tag{4}$$

The SVI and MVI corresponding to problem (1) are defined with the set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and the mapping $F(\mathbf{z}) := (\partial_x f(\mathbf{x}, \mathbf{y}), \partial_y[-f(\mathbf{x}, \mathbf{y})])^\top$ where $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. The main contributions of this paper are summarized below:

- We propose a generic algorithm motivated by the inexact proximal point method (Davis and Grimmer, 2017) for solving a class of non-convex non-concave min-max problems, whose objective function $f(\mathbf{x}, \mathbf{y})$ is weakly convex in $\mathbf{x}$ for any fixed $\mathbf{y} \in \mathcal{Y}$ and weakly concave in $\mathbf{y}$ for any fixed $\mathbf{x} \in \mathcal{X}$. The algorithm consists of approximately solving a sequence of strongly monotone VIs constructed by adding a strongly monotone mapping to $F(\mathbf{z})$ with a sequentially updated proximal center.

- We prove the theoretical convergence of the proposed algorithm under the key assumption that *there exists a solution to the MVI$(F, \mathcal{Z})$ corresponding to (1).* We establish the iteration complexities for finding a nearly ϵ-stationary solution when different algorithms, including the stochastic subgradient method, the gradient descent method, extragradient method, and the variance reduction methods, are employed as a subroutine for solving each strongly monotone VI in the proposed framework. In particular, the iteration complexity is $O(1/\epsilon^6)$ when using the stochastic subgradient method and is $\widetilde{O}(1/\epsilon^2)$ when using the gradient method, or the extragradient method under the additional smoothness assumption of f. [1] Here, the complexity when using Nesterov's accelerated method or the extragradient method improves the one when using the gradient method by a constant factor when a condition number of the original problem is large. Furthermore, if f is of a finite-sum form we can use variance reduction methods to improve the complexity. The achieved complexity results are presented in Table 1.

- Moreover, our algorithms are directly applicable to the more general problem of solving the SVI when F is weakly monotone, and our analysis directly implies the non-asymptotic convergence to a nearly ϵ-accurate solution to the SVI under the condition that a solution to the corresponding MVI exists.

Application in Training GAN. The formulation (1) has broad applications in machine learning, statistics and operations research. Here we present one example in machine learning for training GAN (Goodfellow et al., 2014; Arjovsky et al., 2017; Gulrajani et al., 2017). GAN refers to a powerful class of generative models that cast generative modeling as a game between two networks: a generator network produces synthetic data out of noise and a discriminator network discriminates between the true data and the generator's output. Let us consider WGAN (Arjovsky et al., 2017), a recently proposed variant of GAN, as an example. The min-max problem corresponding to WGAN can be written as

$$\min_{\theta\in\Theta}\max_{\mathbf{w}\in\mathcal{W}}\left\{\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_r}[f_{\mathbf{w}}(\mathbf{x})]-\mathbb{E}_{\mathbf{z}\sim\mathbb{P}_{\mathbf{z}}}[f_{\mathbf{w}}(g_\theta(\mathbf{z}))]\right\},$$

where $f_{\mathbf{w}}(\mathbf{x})$ denotes a Lipschitz continuous function parameterized by $\mathbf{w}$ corresponding to the discriminator, $g_\theta(\mathbf{z})$ denotes the parameterized function corresponding to the generator, $\mathbb{P}_r$ denotes the underlying distribution of the data $\mathbf{x}$, and $\mathbb{P}_{\mathbf{z}}$ denotes the distribution of noise $\mathbf{z}$. Functions $f_{\mathbf{w}}(\cdot)$ and $g_\theta(\cdot)$ are usually represented by deep neural networks. When the deep neural networks induce smooth $f_{\mathbf{w}}(\cdot)$ and $g_\theta(\cdot)$ (for example by using smooth activation functions), it can be showed that the objective function of WGAN is weakly convex in θ and weakly concave in $\mathbf{w}$. More applications of (1) with weakly-convex-weakly-concave function $f(\mathbf{x},\mathbf{y})$ can be found in reinforcement learning (Dai et al., 2018), learning a robust model under heavy-tailed noise (Audibert and Catoni, 2011), adversarial learning (Sinha et al., 2017), etc (cf. examples provided in the Appendix).

2. Related Work

There is a growing interest on first-order algorithms for solving non-convex problems, e.g., (Ghadimi and Lan, 2013; Yang et al., 2016; Ghadimi and Lan, 2016; Reddi et al., 2016b,a; Lan and Yang, 2018; Allen-Zhu, 2017; Allen-Zhu and Hazan, 2016; Davis and Drusvyatskiy, 2018b,a; Zhang and He, 2018; Davis and Drusvyatskiy, 2018a), and this list is by no means complete. The proposed algorithm shares the similarity with several of these previous works (Davis and Grimmer, 2017; Chen et al., 2018; Lan and Yang, 2018; Allen-Zhu, 2017; Chen and Yang, 2018) by using techniques related to the proximal-point method, which has a long history (Rockafellar, 1976). The work of (Shamir, 2020; Zhang et al., 2020) considered first-order algorithms for nonsmooth nonconvex minimization problems. However, their

1. $\widetilde{O}$ suppresses all logarithmic factors.

analysis cannot be directly extended to weakly-convex weakly-concave min-max problems, which is more challenging.

Several recent works (Chen et al., 2017; Qian et al., 2018; Sinha et al., 2017; Rafique et al., 2018; Dai et al., 2018) have considered non-convex min-max problems and their applications in machine learning. However, their algorithms and analysis are either built on restricted assumptions (e.g., the maximization problem can be solved exactly) or only applied to a much smaller family non-convex min-max problems. Rafique et al. (2018) considered weakly-convex and concave min-max problems. Relying on the concavity of the maximization part, they are able to establish the convergence to a nearly stationary point of the function $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. Thekumparampil et al. (2019) developed deterministic first-order algorithms for strongly-convex-concave and nonconvex-concave min-max problems respectively. Lin et al. (2020) provided near-optimal algorithms for strongly-convex-strongly-concave min-max problems. In contrast, the problem considered in this paper is weakly-convex and weakly-concave and covers even more applications in machine learning (e.g., GAN training).

Recently, there emerges a wave of studies that analyze the convergence properties of some algorithms for training GAN (Cherukuri et al., 2017; Daskalakis and Panageas, 2018; Daskalakis et al., 2017; Heusel et al., 2017; Nagarajan and Kolter, 2017; Grnarova et al., 2017). However, their results are either asymptotic (Daskalakis et al., 2017; Heusel et al., 2017; Nagarajan and Kolter, 2017; Cherukuri et al., 2017) or their analysis require strong assumptions of the problem (Nagarajan and Kolter, 2017; Grnarova et al., 2017) (e.g., the problem is concave in maximization). We also notice that two recent papers (Gidel et al., 2018; Mertikopoulos et al., 2018) have also considered the algorithms for min-max problems from the perspective of variational inequalities. However, the analysis in (Gidel et al., 2018) is for convex-concave problems or monotone variational inequalities. Mertikopoulos et al. (2018) proved asymptotic convergence to saddle points of the min-max problem under a strong coherence assumption, i.e., every saddle point is a solution to the corresponding MVI. In contrast, we only assume the MVI has a solution and prove non-asymptotic convergence to a nearly ϵ-stationary point.

Both SVI and MVI have a long history in the literature (Harker and Pang, 1990). When the set-valued mapping F is non-monotone, there have been many studies that design and analyze algorithms for finding a solution to the SVI problem (Solodov and Svaiter, 1999; Bao and Khanh, 2005; Burachik and Millan, 2016; Allevi et al., 2006; Dang and Lan, 2015; Iusem et al., 2017). However, the main difference between these works and the present work is that their convergence analysis is asymptotic except for (Dang and Lan, 2015; Iusem et al., 2017). Dang and Lan (2015) established the first non-asymptotic convergence for non-monotone SVI by deterministic algorithms when the mapping $F(\mathbf{z})$ is single-valued and Lipschitz (or Hölder) continuous. Iusem et al. (2017) analyzed a variant of stochastic extragradient method with increasing mini-batch size for solving SVI with a single-valued and Lipschitz continuous mapping and obtained several complexity results but for a different convergence measure from (Dang and Lan, 2015) and ours. In contrast, our work provides the first non-asymptotic convergence of stochastic and deterministic algorithms for solving the SVI problems with a set-valued mapping that is non-Lipschitz and non-Hölder continuous but weakly monotone.

3. Preliminaries

We present some preliminaries in this section. For simplicity, we consider (1) defined in the Euclidean space with inner product $\langle \mathbf{z}, \mathbf{z}' \rangle = \mathbf{z}^\top \mathbf{z}'$. We use $\|\cdot\|$ to represent the Euclidean norm. Let $\text{Proj}_{\mathcal{Z}}[\mathbf{z}]$ denote an Euclidean projection mapping that projects $\mathbf{z}$ onto the set $\mathcal{Z}$. Given a function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we define the *(Fréchet) subdifferential* of h as

$$\partial h(\mathbf{x}) = \left\{ \boldsymbol{\zeta} \in \mathbb{R}^d \middle| \begin{array}{l} h(\mathbf{x}') \geq h(\mathbf{x}) + \boldsymbol{\zeta}^\top(\mathbf{x}' - \mathbf{x}) + o(\|\mathbf{x}' - \mathbf{x}\|), \\ \mathbf{x}' \to \mathbf{x} \end{array} \right\}, \tag{5}$$

where each element in $\partial h(\mathbf{x})$ is called a *(Fréchet) subgradient* of h at $\mathbf{x}$. In this paper, we will analyze the convergence of an iterative algorithm for solving (1) through the lens of variational inequalities. To this end, we first introduce some background related to variational inequalities.

Definition 1 *A set-valued mapping $F(\mathbf{z}) : \mathcal{Z} \rightrightarrows \mathbb{R}^d$ is said to be monotone if $\langle \boldsymbol{\xi} - \boldsymbol{\xi}', \mathbf{z} - \mathbf{z}' \rangle \geq 0$, μ-strongly monotone for $\mu > 0$ if $\langle \boldsymbol{\xi} - \boldsymbol{\xi}', \mathbf{z} - \mathbf{z}' \rangle \geq \mu \|\mathbf{z} - \mathbf{z}'\|^2$, and ρ-weakly monotone for $\rho > 0$ if $\langle \boldsymbol{\xi} - \boldsymbol{\xi}', \mathbf{z} - \mathbf{z}' \rangle \geq -\rho \|\mathbf{z} - \mathbf{z}'\|^2$, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, $\boldsymbol{\xi} \in F(\mathbf{z})$, and $\boldsymbol{\xi}' \in F(\mathbf{z}')$.*

By a slight abuse of notation, when $F(\mathbf{z})$ is a singleton set, we will use $F(\mathbf{z})$ to represent the single element in the set. A (single-valued) mapping $F(\mathbf{z}) : \mathcal{Z} \to \mathbb{R}^d$ is said to be L-Lipschitz continuous if $\|F(\mathbf{z}) - F(\mathbf{z}')\| \leq L\|\mathbf{z} - \mathbf{z}'\|$, $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}$. It is well-known that SVI$(F, \mathcal{Z})$ has a *unique solution* if F is μ-strongly monotone (Nesterov and Scrimali, 2011).

For the min-max problem (1), we define $\mathbf{z} = (\mathbf{x}, \mathbf{y})^\top$, $F(\mathbf{z}) \equiv (\partial_x f(\mathbf{x}, \mathbf{y}), -\partial_y f(\mathbf{x}, \mathbf{y}))^\top$ and $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$. The SVI corresponding to (1) is defined by such a $F(\mathbf{z})$ and $\mathcal{Z}$. If $F(\mathbf{z})$ is L-Lipchitz continuous it is also L-weakly monotone. However, Lipchitz continuity of F is not necessary for F to be weakly monotone. Below, we show that if $f(\mathbf{x}, \mathbf{y})$ is weakly convex in terms of $\mathbf{x}$ and weakly concave in terms of $\mathbf{y}$, then its corresponding F is weakly monotone. To this end, we first introduce the definition of weakly convex and weakly concave.

Definition 2 *$f(\mathbf{x}, \mathbf{y})$ is ρ-weakly-convex-weakly-concave if for any $\mathbf{y} \in \mathcal{Y}$, $f(\mathbf{x}, \mathbf{y}) + \frac{\rho}{2}\|\mathbf{x}\|^2$ is convex in $\mathbf{x}$, and, for any $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}, \mathbf{y}) - \frac{\rho}{2}\|\mathbf{y}\|^2$ is concave in $\mathbf{y}$.*

In Section B in Appendix, we present some examples of the min-max problem whose objective function is weakly-convex and weakly-concave, which are not necessarily smooth functions. The following lemma shows the connection between weak-convexity weak-concavity and weak-monotonicity. Its proof and all missing proofs are included in the Appendix.

Lemma 3 *$f(\mathbf{x}, \mathbf{y})$ is ρ-weakly-convex-weakly-concave if and only if $F(\mathbf{z})$ is ρ-weakly monotone.*

The following lemma is standard but critical for obtaining our results. The counterpart of this lemma for minimization problems can be found in (Davis and Grimmer, 2017).

Lemma 4 *If $F(\mathbf{z}) : \mathcal{Z} \rightrightarrows \mathbb{R}^d$ is ρ-weakly monotone, the mapping $F^{\gamma}_{\mathbf{w}}(\mathbf{z}) \equiv F(\mathbf{z}) + \frac{1}{\gamma}(\mathbf{z} - \mathbf{w})$ is $(\frac{1}{\gamma} - \rho)$-strongly monotone on $\mathcal{Z}$ for any $0 < \gamma < \rho^{-1}$ and any $\mathbf{w} \in \mathcal{Z}$.*

A point $\mathbf{z} \in \mathcal{Z}$ is called first-order stationary point of (1) if $\mathbf{z} \in \mathcal{F}_*$ with $\mathcal{F}_*$ defined in (2). An iterative algorithm typically only finds an ϵ-stationary solution which is a solution $\mathbf{z} \in \mathcal{Z}$ such that $\text{dist}(\mathbf{0}, \partial(f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Z}}(\mathbf{x}, \mathbf{y})) \le \epsilon$, where $\partial(f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Z}}(\mathbf{x}, \mathbf{y})) \equiv \partial_x[f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{X}}(\mathbf{x})] \times \partial_y[-f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Y}}(\mathbf{y})]$ and $\text{dist}(\mathbf{z}, \mathcal{S})$ denotes the Euclidean distance from a point to a set $\mathcal{S}$. The non-smoothness nature of the problem makes it challenging to find an ϵ-stationary solution for an iterative algorithm. For example, consider the convex minimization $\min_{z\in[-1,1]} |z|$ and has a solution at 0. Hence, if $\bar{z}$ is very close to 0 but not 0, we always have $|\bar{\xi}| = 1$ for any $\bar{\xi} \in \partial|\bar{z}|$. Hence, we consider the following notion of a nearly stationary point for a non-smooth min-max problem.

Definition 5 *A point $\mathbf{w} \in \mathcal{Z}$ is called a nearly ϵ-stationary solution to (1) if there exists $\bar{\mathbf{w}} = (\bar{\mathbf{u}}, \bar{\mathbf{v}})^\top \in \mathcal{Z}$ and $c > 0$ such that $\|\mathbf{w} - \bar{\mathbf{w}}\| \le c\epsilon$, and $dist(0, \partial(f(\bar{\mathbf{u}}, \bar{\mathbf{v}}) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}, \bar{\mathbf{v}}))] \le \epsilon$.*

Such a notion of nearly stationary has been utilized in several works for tackling non-smooth non-convex minimization problems (Davis and Drusvyatskiy, 2018b; Davis and Grimmer, 2017; Chen et al., 2018). In order to show the existence of a nearly ϵ-stationary solution, we define a *proximal-point mapping* of F at a proximal center $\mathbf{w}$ as

$$F_{\mathbf{w}}^{\gamma}(\mathbf{z}) \equiv F(\mathbf{z}) + \frac{1}{\gamma}(\mathbf{z} - \mathbf{w}) \tag{6}$$

where $0 < \gamma < \rho^{-1}$. According to Lemma 4, $F_{\mathbf{w}}^{\gamma}$ is $(\frac{1}{\gamma} - \rho)$-strongly monotone so that $\text{SVI}(F_{\mathbf{w}}^{\gamma}, \mathcal{Z})$ has a unique solution denoted by $\bar{\mathbf{w}}$. The following lemma is a foundation of our algorithm and analysis, which shows that as long as we find a solution $\mathbf{w}$ such that $\|\mathbf{w} - \bar{\mathbf{w}}\| \le \gamma\epsilon$, it is a nearly ϵ-stationary solution.

Lemma 6 *Let $F_{\mathbf{w}}^{\gamma}$ be defined in (6) for $0 < \gamma < \rho^{-1}$ and $\mathbf{w} = (\mathbf{u}, \mathbf{v})^\top \in \mathcal{Z}$. Denote by $\bar{\mathbf{w}} = (\bar{\mathbf{u}}, \bar{\mathbf{v}})^\top$ the solution to $SVI(F_{\mathbf{w}}^{\gamma}, \mathcal{Z})$. We have*

$$dist(0, \partial(f(\bar{\mathbf{u}}, \bar{\mathbf{v}}) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}, \bar{\mathbf{v}})) \le \|\mathbf{w} - \bar{\mathbf{w}}\|/\gamma. \tag{7}$$

Before ending this section, we formally present the basic assumptions used in our analysis.

Assumption 1 *(i) The set $\mathcal{Z}$ is convex and compact so that there exists $D > 0$ such that $\max_{\mathbf{z},\mathbf{z}'\in\mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\| \le D$. (ii) The mapping F is ρ-weakly monotone. (iii) The $MVI(F, \mathcal{Z})$ problem has a solution.*

It is notable that the last assumption has been used by most previous works for solving non-monotone SVI (Solodov and Svaiter, 1999; Wang et al., 2001; Bao and Khanh, 2005; Burachik and Millan, 2016; Dang and Lan, 2015; Iusem et al., 2017). It can be shown that when $F(\mathbf{z})$ satisfies some generalized notion of monotonicity (e.g., pseudomonotone, quasi-monotone), a solution of $\text{MVI}(F, \mathcal{Z})$ exists. Indeed, a similar assumption (for non-convex minimization) has been made for analyzing the convergence of stochastic gradient descent for learning neural networks (Li and Yuan, 2017) and also was observed in practice (Kleinberg et al., 2018).

Algorithm 1 Inexact Proximal Point (IPP) Method for Weakly-Monotone SVI

1: **Input:** integer $K \geq 1$, step size $\eta_k > 0$, integer $T_k \geq 1$, weight $\theta_k > 0$ non-decreasing in k, and $0 < \gamma < \rho^{-1}$
2: **for** $k = 0, \ldots, K-1$ **do**
3: Let $F_k \equiv F^\gamma_{\mathbf{z}_k} = F(\mathbf{z}) + \gamma^{-1}(\mathbf{z} - \mathbf{z}_k)$
4: $\mathbf{z}_{k+1} = \text{ApproxSVI}(F_k, \mathcal{Z}, \mathbf{z}_k, \eta_k, T_k)$
5: **end for**
6: **Output:** $\mathbf{z}_\tau$, where τ is randomly from $\{0, 1, \ldots, K-1\}$ with $\text{Prob}(\tau = k) = \frac{\theta_k}{\sum_{l=0}^{K-1}\theta_l}$.

4. A Generic Algorithmic Framework with a General Convergence Result

In this section, we will present a generic algorithmic framework for solving the saddle-point problem (1) through the lens of VI. The method we propose is called the inexact proximal point (IPP) method. This method consists of solving a sequence of strongly monotone SVIs defined by the mappings $F^\gamma_{\mathbf{z}_k}$ in (6) with a sequentially updated proximal center $\mathbf{z}_k$. In particular, an appropriate first-order algorithm is employed to find an approximate solution $\mathbf{z}_{k+1}$ to $\text{SVI}(F^\gamma_{\mathbf{z}_k}, \mathcal{Z})$, which is then used as the proximal center that defines $\text{SVI}(F^\gamma_{\mathbf{z}_{k+1}}, \mathcal{Z})$. This method is described in Algorithm 1. A subroutine $\text{ApproxSVI}(F_k, \mathcal{Z}, \mathbf{z}_k, \eta_k, T_k)$ is called to approximately solve $\text{SVI}(F_k, \mathcal{Z})$, where $\mathbf{z}_k$ is the initial solution of the subroutine, η_k denotes the step size and T_k denotes the number of iterations performed in the subroutine.

The strength of our framework is that it decomposes a complex non-convex non-concave minmax (or non-monotone SVI) problem into a sequence of easier strongly-convex strongly-concave (or strongly monotone SVI) problems. Hence, it allows one to leverage algorithms and convergence theory for strongly-convex strongly-concave (or strongly monotone SVI) problems to analyze the convergence for the original problem. Such approach is the first time used for solving non-convex non-concave minmax problems and our analysis is novel. A generic convergence result of Algorithm 1 conditioned on a particular sequence of $\mathbf{z}_k$ generated by ApproxSVI is stated below.

Theorem 7 *Suppose Assumption 1 holds, and ApproxSVI$(F_k, \mathcal{Z}, \mathbf{z}_k, \eta_k, T_k)$ in Algorithm 1 ensures*

$$\max_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}[\xi_{k+1}^\top(\mathbf{z}_{k+1} - \mathbf{z})|\mathbf{z}_k] \leq \frac{c}{k+1}, \tag{8}$$

for $k = 0, 1, \ldots, K-1$, where $c > 0$, $\xi_{k+1} \in F_k(\mathbf{z}_{k+1})$ and $\mathbb{E}[\cdot|\mathbf{z}_k]$ is the conditional expectation conditioning on $\mathbf{z}_k$. By choosing $\gamma = \frac{1}{2\rho}$ and $\theta_k = (k+1)^\alpha$ with $\alpha \geq 1$ in Algorithm 1, we have

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \frac{2D^2(\alpha+1)}{K} + \frac{4c(\alpha+1)}{K\rho}. \tag{9}$$

where $\bar{\mathbf{z}}_\tau$ is the unique solution to SVI$(F^\gamma_{\mathbf{z}_\tau}, \mathcal{Z})$.

Remark: The total complexity of Algorithm 1 for finding a nearly ϵ-stationary solution depends on the complexity of ApproxSVI for computing each $\mathbf{z}_{k+1}$ satisfying (8), which corresponds to solving $\text{SVI}(F^\gamma_{\mathbf{z}_k}, \mathcal{Z})$ approximately with an increasing accuracy as k increases.

Algorithm 2 Stochastic Subgradient Method for SVI($F, \mathcal{Z}$): SG($F, \mathcal{Z}, \mathbf{z}^{(0)}, \eta, T$)

1: **Input:** mapping F, set $\mathcal{Z}$, $\mathbf{z}^{(0)} \in \mathcal{Z}$, $\eta > 0$, $T \geq 1$
2: **for** $t = 0, ..., T-1$ **do**
3: $\quad \mathbf{z}^{(t+1)} = \text{Proj}_{\mathcal{Z}}\left(\mathbf{z}^{(t)} - \eta\boldsymbol{\zeta}(\mathbf{z}^{(t)})\right)$ where $\boldsymbol{\zeta}(\mathbf{z}^{(t)})$ satisfies $\mathbb{E}[\boldsymbol{\zeta}(\mathbf{z}^{(t)})] \in F(\mathbf{z}^{(t)})$.
4: **end for**
5: **Output:** $\mathbf{z}^{(\tau)}$, where τ is uniformly randomly from $\{0, 1, \ldots, T-1\}$.

Below, we present several candidates for ApproxSVI with their convergence properties. We first consider (stochastic) subgradient method without imposing Lipchitz continuity assumption of F. In next section, we derive improved rates when F is Lipchitz continuous. With a specific algorithm $\mathcal{A}$ for solving SVI($F^{\gamma}_{\mathbf{z}_k}, \mathcal{Z}$) at each stage of Algorithm 1, the resulting algorithm is named as IPP-$\mathcal{A}$.

Proposition 8 summarizes the convergence result of stochastic subgradient method for solving SVI($F_k, \mathcal{Z}$) approximately, and Corollary 9 is a corollary of Theorem 9 that states the convergence result of IPP-SG. Here, we only show stochastic subgradient method in Algorithm 2.

Proposition 8 *Suppose F_k is monotone, $\mathbb{E}[\boldsymbol{\zeta}(\mathbf{z})] \in F_k(\mathbf{z})$, and $\mathbb{E}\|\boldsymbol{\zeta}(\mathbf{z})\|^2 \leq G_k^2$ for any $\mathbf{z} \in \mathcal{Z}$. Algorithm 2 applied to SVI($F_k, \mathcal{Z}$) guarantees that*

$$\max_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}[(\boldsymbol{\xi}^{(\tau)})^\top(\mathbf{z}^{(\tau)} - \mathbf{z})] \leq \frac{D^2}{2\eta T} + \frac{\eta G_k^2}{2}, \text{ where } \boldsymbol{\xi}^{(\tau)} \in F_k(\mathbf{z}^{(\tau)}). \tag{10}$$

Remark: We would like to emphasize that as long as an algorithm can find a $\mathbf{z}^{(\tau)}$ such that $\max_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}[(\boldsymbol{\xi}^{(\tau)})^\top(\mathbf{z}^{(\tau)} - \mathbf{z})] \leq O(1/\sqrt{T})$ in T iterations, it can be used as ApproxSVI and IPP will achieve the same total iteration as when the stochastic subgradient method is used. One example that is of particular interest to the deep learning community (e.g., for training GAN (Gulrajani et al., 2017)) is the Adam-style stochastic algorithm (Kingma and Ba, 2015; Reddi et al., 2018).

Corollary 9 *For problem (1), assume that Assumption 1 holds, and for any $\mathbf{z} \in \mathcal{Z}$ there exists $\boldsymbol{\zeta}(\mathbf{z}) \in F(\mathbf{z})$ such that $\mathbb{E}[\boldsymbol{\zeta}(\mathbf{z})] \in F(\mathbf{z})$, and $\mathbb{E}[\|\boldsymbol{\zeta}(\mathbf{z})\|^2] \leq G^2$. Suppose Algorithm 2 is used as ApproxSVI with $\eta_k = \frac{D}{G_k(k+1)}$, $G_k = G + 2\rho D$ and $T_k = (k+1)^2$. By choosing $\gamma = \frac{1}{2\rho}$, $\theta_k = (k+1)^{\alpha}$ with $\alpha \geq 1$, and $K = \frac{(16\rho^2 D^2 + 4\rho DG)(\alpha+1)}{\epsilon^2}$ in Alg. 1, we have*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \gamma^2\epsilon^2, \quad \mathbb{E}[dist^2(0, \partial(f(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau))] \leq \epsilon^2$$

where $\bar{\mathbf{z}}_\tau = (\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau)^\top$ is the solution to SVI($F^{\gamma}_{\mathbf{z}_\tau}, \mathcal{Z}$) The total iteration complexity of $O\left(\frac{\max(\rho^6,\rho^3)}{\epsilon^6}\right)$.

Remark: To the best of our knowledge, this is **the first non-asymptotic convergence** of stochastic subgradient algorithms for solving non-convex non-concave problems.

5. Improved Rates for Lipchitz Continuous Operator

In this section, we present improved rates when F is single-valued and Lipchitz continuous. In particular, we consider gradient method, extragradient method and variance reduction methods for finite-sum problems. The key to the improved rates is that these methods could achieve faster convergence rates for strongly monotone problems SVI$(F_k, \mathcal{Z})$, which have been analyzed in the literature (Nesterov and Scrimali, 2011; Rafique et al., 2018; Palaniappan and Bach, 2016; Gidel et al., 2018). However, there is still a gap between the existing convergence results for strongly monotone problems and the requirement (8) in Theorem 1 because the existing faster convergence for strongly monotone problems are for the distance to the optimal solution. The following lemma can bridge the gap that allows us to use existing faster convergence results for Lipchitz continuous operator.

Lemma 10 *Suppose that there exists an algorithm for a monotone SVI$(F, \mathcal{Z})$ with L-Lipschitz continuous single-valued mapping $F(\mathbf{z})$ that returns a solution $\widehat{\mathbf{z}}$. Let $\mathbf{w}_*$ be the solution of SVI$(F, \mathcal{Z})$. Then by constructing $\bar{\mathbf{z}} = Proj_{\mathcal{Z}}(\widehat{\mathbf{z}} - \eta F(\widehat{\mathbf{z}}))$ with $\eta = 1/(\sqrt{2}L)$, we have* $\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z}) \leq DL(2+\sqrt{2})\|\widehat{\mathbf{z}} - \mathbf{w}_*\|$.

5.1 Using Gradient Descent/ Extragradient Method

The gradient descent method is presented in Algorithm 3 and the extragradient method is presented in Algorithm 4 and their linear convergence for the gap $\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z})$ is stated in the following proposition. It is worth mentioning that the last step $\bar{\mathbf{z}} = \text{Proj}_{\mathcal{Z}}(\mathbf{w}^{(T+1)} - F(\mathbf{w}^{(T+1)})/(\sqrt{2}L))$ in Algorithm 4 following Lemma 10 is important to prove the linear convergence of the gap $\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z})$.

Proposition 11 *Suppose that F_k is single-valued and L_k-Lipschitz continuous and is μ-strongly monotone. Algorithm 3 applied to SVI$(F_k, \mathcal{Z})$ with $\eta = \frac{\mu}{2L_k^2}$ guarantees*

$$\max_{\mathbf{z}\in\mathcal{Z}} F_k(\mathbf{z}^{(T)})^\top(\mathbf{z}^{(T)} - \mathbf{z}) \leq 4D^2 L_k \beta \exp(-\frac{T-1}{4\beta^2}), \text{ where } \beta = L_k/\mu \geq 1.$$

Algorithm 4 applied to SVI$(F_k, \mathcal{Z})$ with $\eta = 1/(4L_k)$ guarantees that for any $\mathbf{z} \in \mathcal{Z}$

$$\max_{\mathbf{z}\in\mathcal{Z}} F_k(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z}) \leq 4D^2 L_k \exp(-T/(8\beta)), \text{ where } \beta = L_k/\mu \geq 1.$$

The following two corollaries summarize the convergence results of IPP-GD and IPP-EG, respectively.

Corollary 12 *For problem (1), assume that Assumption 1 holds and F is single-valued and L-Lipschitz continuous. Suppose Algorithm 3 is used as ApproxSVI with $\eta_k = \frac{\rho}{2(L+2\rho)^2}, T_k = 1 + \frac{4(L+2\rho)^2}{\rho^2}\log(\frac{8(L+2\rho)^2(k+1)}{\rho^2})$. By choosing $\gamma = 1/(2\rho)$, $\theta_k = (k+1)^\alpha$ with $\alpha \geq 1$, and $K = \frac{16\rho^2 D^2(\alpha+1)}{\epsilon^2}$ in Algorithm 1, we have*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \gamma^2\epsilon^2, \quad \mathbb{E}[dist^2(0, \partial(f(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau))] \leq \epsilon^2$$

where $\bar{\mathbf{z}}_\tau = (\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau)^\top$ is the solution to SVI$(F^\gamma_{\mathbf{z}_\tau}, \mathcal{Z})$. The total iteration complexity of $O(\log(\frac{1}{\epsilon})\frac{L^2}{\epsilon^2})$.

Algorithm 3 GD$(F, \mathcal{Z}, \mathbf{z}^{(0)}, \eta, T)$

for $t = 0, ..., T-1$ **do**
  $\mathbf{z}^{(t+1)} = \text{Proj}_{\mathcal{Z}}\left(\mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)})\right)$
end for
Output: $\mathbf{z}^{(T)}$

Algorithm 4 EG$(F, \mathcal{Z}, \mathbf{w}^{(0)}, \eta, T)$

for $t = 0, ..., T$ **do**
  $\mathbf{z}^{(t)} = \text{Proj}_{\mathcal{Z}}\left(\mathbf{w}^{(t)} - \eta F(\mathbf{w}^{(t)})\right)$
  $\mathbf{w}^{(t+1)} = \text{Proj}_{\mathcal{Z}}\left(\mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t)})\right)$
end for
Output: $\bar{\mathbf{z}} = \text{Proj}_{\mathcal{Z}}(\mathbf{w}^{(T+1)} - F(\mathbf{w}^{(T+1)})/(\sqrt{2}L))$

Corollary 13 *Under the same assumption in Corollary 12, Algorithm 1 with* $\gamma = 1/(2\rho)$, $\theta_k = (k+1)^\alpha$ *with* $\alpha > 1$, $T_k = \frac{8(L+2\rho)}{\rho}\log(\frac{4(k+1)(L+2\rho)}{\rho})$, *and* $K = 24\rho^2 D^2(\alpha+1)/\epsilon^2$ *guarantees*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \gamma^2\epsilon^2, \quad E[dist^2(0, \partial(f(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau))] \leq \epsilon^2 \tag{11}$$

where $\bar{\mathbf{z}}_\tau$ *is the solution to SVI*$(F^\gamma_{\mathbf{z}_\tau}, \mathcal{Z})$*. The total iteration complexity is* $O(\log(\frac{1}{\epsilon})\frac{L\rho}{\epsilon^2})$.

Remark: The improvement of using the extragradient method over the gradient descent method lies at the better dependence on the condition number $L/\rho \geq 1$. In particular, IPP-GD has a complexity of $\widetilde{O}((L/\rho)^2\rho^2/\epsilon^2)$ and IPP-EG has a complexity of $\widetilde{O}((L/\rho)\rho^2/\epsilon^2)$.

5.2 Using Variance Reduction methods

Next, we discuss how to use variance reduction methods for solving each strongly monotone SVI when the underlying $F(\mathbf{z})$ is μ-strongly monotone, has a finite-sum form

$$F(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^n F_i(\mathbf{z}),$$

where each F_i is L-Lipschitz continuous. Several studies have considered variance reduction algorithms for solving strongly monotone SVI or the strongly convex and strongly concave min-max problems (Palaniappan and Bach, 2016; Rafique et al., 2018). It is notable that in these studies, linear convergence is only proved for the point convergence, i.e., $\mathbb{E}[\|\widehat{\mathbf{z}} - \mathbf{w}_*\|]$ where $\mathbf{w}_*$ denotes the solution of the strongly monotone SVI. However, by using an additional gradient update as in Lemma 10, we can get the linear convergence for $\mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} F_k(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z})]$. Combining this result, existing convergence results of variance reduction algorithms for solving each strongly monotone SVI$(F_k, \mathcal{Z})$ and the result in Theorem 7, we can derive the complexity for solving the original SVI$(F, \mathcal{Z})$ or the min-max saddle-point problem. According to (Palaniappan and Bach, 2016), the complexity of SVRG-based algorithm for solving a strongly convex and strongly concave min-max smooth problem in the sense that $\mathbb{E}[\|\widehat{\mathbf{z}} - \mathbf{w}_*\|] \leq \epsilon$ is $O((n + L^2/\mu^2)\log(1/\epsilon))$, which gives a total gradient complexity of $\widetilde{O}((n + L^2/\rho^2)/\epsilon^2)$ for finding a solution $\mathbf{z}_\tau$ that satisfies $\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|] \leq \epsilon^2$ where $\bar{\mathbf{z}}_\tau = (\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau)^\top$ is the solution to SVI$(F^\gamma_{\mathbf{z}_\tau}, \mathcal{Z})$. Equivalently, for finding a nearly ϵ-stationary solution of the original problem, the total gradient complexity is $\widetilde{O}((n\rho^2 + L^2)/\epsilon^2)$.

Now we present the details about the approach. Following the setting in (Palaniappan and Bach, 2016), we first equivalently view the SVI problem as the problem of finding a $\mathbf{z}$ such that

$$\mathbf{0} \in F(\mathbf{z}) + \partial\mathbf{1}_{\mathcal{Z}}(\mathbf{z}),$$

Algorithm 5 SVRG Method for SVI(F, Z): SVRG($F, \mathcal{Z}, \mathbf{w}^{(0)}, T$)

Input: $\mu-$Strongly Monotone and L-Lipschitz continuous mapping $F = \frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{z})$, set $\mathcal{Z}$, $\mathbf{w}^{(0)} \in \mathcal{Z}$, and an integer $T \geq 1$.
Initialize $\eta = \mu/(4(L+\mu)^2)$, $S = \lceil 4\log(4)(L+\mu)^2/\mu^2 \rceil$, $A(\mathbf{z}) = \mu(\mathbf{z}-\mathbf{w}) + \partial \mathbf{1}_{\mathcal{Z}}(\mathbf{z})$, $B(\mathbf{z}) = F(\mathbf{z}) - \mu(\mathbf{z}-\mathbf{w})$, and $B_i(\mathbf{z}) = \frac{1}{n}F_i(\mathbf{z}) - \frac{\mu}{n}(\mathbf{z}-\mathbf{w})$ for $i = 1, \dots, n$.
for $t = 0, ..., T-1$ **do**
  $\bar{B}^{(t)} = B(\mathbf{w}^{(t)})$ and $\mathbf{z}^{(0)} = \mathbf{w}^{(t)}$
  for $s = 0, ..., S-1$ **do**
    Sample i uniformly randomly from $\{1, 2, \dots, n\}$.
    $B^{(s)} = \bar{B}^{(t)} - nB_i(\mathbf{w}^{(t)}) + nB_i(\mathbf{z}^{(s)})$
    $\mathbf{z}^{(s+1)} = (I + \eta A)^{-1}(\mathbf{z}^{(s)} - \eta B^{(s)})$
  end for
  $\mathbf{w}^{(t+1)} = \mathbf{z}^{(S)}$
end for
Output: $\bar{\mathbf{z}} = \mathrm{Proj}_{\mathcal{Z}}\left(\mathbf{w}^{(T)} - F(\mathbf{w}^{(T)})/(\sqrt{2}L)\right)$

where $\partial \mathbf{1}_{\mathcal{Z}}(\mathbf{z}) = N_{\mathcal{Z}}(\mathbf{z})$ is the normal cone of $\mathcal{Z}$ at $\mathbf{z}$. Note that $F(\mathbf{z}) + \partial \mathbf{1}_{\mathcal{Z}}(\mathbf{z})$ is still a strongly monotone mapping. Furthermore, we reformulate $F(\mathbf{z}) + \partial \mathbf{1}_{\mathcal{Z}}(\mathbf{z})$ as

$$F(\mathbf{z}) + \partial \mathbf{1}_{\mathcal{Z}}(\mathbf{z}) = A(\mathbf{z}) + B(\mathbf{z}) = A(\mathbf{z}) + \sum_{i=1}^{n} B_i(\mathbf{z}),$$

where $A(\mathbf{z}) := \mu(\mathbf{z}-\mathbf{w}) + \partial \mathbf{1}_{\mathcal{Z}}(\mathbf{z})$ is μ-strongly monotone, $B(\mathbf{z}) := F(\mathbf{z}) - \mu(\mathbf{z}-\mathbf{w})$ is monotone and $(L+\mu)$-Lipschitz-continuous, and $B_i(\mathbf{z}) := \frac{1}{n}F_i(\mathbf{z}) - \frac{\mu}{n}(\mathbf{z}-\mathbf{w})$ is $(L+\mu)/n$-Lipschitz-continuous for $i = 1, \dots, n$. Based on this structure, we present the SVRG algorithm for SVI by Palaniappan and Bach (2016) in Algorithm 5. Note that Algorithm 5 utilizes the resolvent operator of the mapping A defined above, which is the mapping $(I+\eta A)^{-1}(\mathbf{z}) = \arg\min_{\mathbf{z}' \in \mathcal{Z}} \frac{1}{2\eta}\|\mathbf{z}'-\mathbf{z}\|^2 + \frac{\mu}{2}\|\mathbf{z}'-\mathbf{w}\|^2$.

Proposition 14 *When $F(\mathbf{z})$ is single-valued, μ-strongly monotone, and of the form $F = \frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{z})$ with each $F_i(\mathbf{z})$ L-Lipschitz continuous, Algorithm 5 guarantees that for any $\mathbf{z} \in \mathcal{Z}$*

$$\mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z})] \leq D^2 L(2+\sqrt{2})(\sqrt{3}/2)^T. \tag{12}$$

In addition, we have

$$\mu\|\bar{\mathbf{z}} - \mathbf{w}_*\|^2 \leq D^2 L(2+\sqrt{2})(\sqrt{3}/2)^T$$

where $\mathbf{w}_*$ denotes a solution to SVI($F, \mathcal{Z}$).

Proof According to Section D.1 in (Palaniappan and Bach, 2016), with the given S and η, Algorithm 5 guarantees that

$$\mathbb{E}\|\mathbf{w}^{(T)} - \mathbf{w}_*\|^2 \leq \left(\frac{3}{4}\right)^T \|\mathbf{w}^{(0)} - \mathbf{w}_*\|^2,$$

where $\mathbf{w}_*$ denotes a solution to SVI$(F, \mathcal{Z})$ (so that $\mathbf{0} \in F(\mathbf{w}_*) + \partial \mathbf{1}_{\mathcal{Z}}(\mathbf{w}_*)$). Following Lemma 10, we have

$$\mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z})] \leq DL(2+\sqrt{2})\mathbb{E}\|\mathbf{w}^{(T)} - \mathbf{w}_*\| \leq D^2L(2+\sqrt{2})(\sqrt{3}/2)^T$$

∎

Note that the complexity of Algorithm 5 for T iterations is $O(T(n + L^2/\mu^2))$.

Corollary 15 *Suppose Assumption 1 holds and $F(\mathbf{z})$ is single-valued and of the form $F = \frac{1}{n}\sum_{i=1}^n F_i(\mathbf{z})$ with each $F_i(\mathbf{z})$ L-Lipschitz continuous, and Algorithm 5 is used to implement ApproxSVI. Algorithm 1 with $\gamma = 1/(2\rho)$, $\theta_k = (k+1)^\alpha$ with $\alpha > 1$, $T_k = (1-\sqrt{3}/2)^{-1}\log((2+\sqrt{2})L(k+1)/\rho)$, and a total of stages $K = 6D^2(\alpha+1)/\epsilon^2$ guarantees*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \epsilon^2, \tag{13}$$

and

$$\mathbb{E}[dist^2(0, \partial(f(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau))] \leq \epsilon^2/\gamma^2, \tag{14}$$

where $\bar{\mathbf{z}}_\tau$ is the solution to SVI$(F^\gamma_{\mathbf{z}_\tau}, \mathcal{Z})$. The total iteration complexity is $\widetilde{O}(D^2(n+L^2/\rho^2)/(\epsilon^2))$.

6. Experimental Results

6.1 Synthetic Experiment

We conduct two synthetic experiments to validate the presented theoretical results.

The first experiment focuses on the comparison between IPP-GD and IPP-EG. To this end, we consider $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{x}^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top A\mathbf{y}$, $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_2 \leq x_1 \leq 1\}$, $\mathcal{Y} = \{(y_1, y_2) \in \mathbb{R}^2 : 0 \leq y_2 \leq y_1 \leq 1\}$, where $A = \text{diag}(1, -\rho)$, $0 < \rho < 1$. It is not difficult to show that Assumption 1 holds, i.e., $D = \sqrt{2}$, the mapping $F(\mathbf{x}, \mathbf{y}) = (\partial_{\mathbf{x}} f(\mathbf{x}, y), -\partial_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))^\top$ is ρ-weakly monotone and the corresponding MVI problem has a solution $(0, 0, 0, 0)$. In addition, $F(\mathbf{x}, \mathbf{y})$ is Lipschitz continuous with modulus $L = 1$. For both algorithms, we start from $(0.2, 0.1, 0.2, 0.1)$, set the hyper-parameters of two algorithms according to Corollary 12 and 13 respectively, run both algorithms until reaching a point whose gradient's magnitude is less than 10^{-2}, and report the number of gradient evaluations in Table 6.1. To illustrate the effect of condition number on the convergence, we set different values of ρ in a range $\{10^{-7:1:-1}\}$. From Table 6.1, we can see that both IPP-GD and IPP-EG converge, and IPP-EG converges much faster. It is worth mentioning that when ρ is sufficiently small (i.e. the condition number L/ρ is sufficiently large), IPP-EG is L/ρ times faster than IPP-GD, which is consistent with our theory.

In the second experiment, We aim to verify the theory of IPP-SGD and IPP-SVRG. We consider the following finite sum min-max optimization problem.

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^n \xi_i \left[\frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{x}^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top A\mathbf{y}\right],$$

where $\xi_i \sim N(1, 5)$, $i = 1, \ldots, n-1$, $n = 1000$, and $\xi_n = n - \sum_{i=1}^{n-1}\xi_i$, $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_2 \leq x_1 \leq 1\}$, $\mathcal{Y} = \{(y_1, y_2) \in \mathbb{R}^2 : 0 \leq y_2 \leq y_1 \leq 1\}$, $A = \text{diag}(1, -\rho)$,

ρ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
IPP-GD	9239	500334	14857	144579	1441992	14416139	$> 10^8$
IPP-EG	1854	10942	62	60	60	60	60

Table 2: IPP-EG vs IPP-GD for the synthetic experiment. The numbers indicate the number of gradient evaluations to achieve a point where the gradient's norm is less than 10^{-2}.

ρ ($\epsilon = 10^{-10}$)	10^{-1}	10^{-2}	10^{-3}
IPP-SGD	12710	3051	2515
IPP-SVRG	1002	1006	1503

ρ ($\epsilon = 10^{-1}$)	10^{-1}	10^{-2}	10^{-3}
IPP-SGD	172	70	67
IPP-SVRG	1002	1002	1025

Table 3: IPP-SGD vs IPP-SVRG for the synthetic experiment. The left (right) table is the number of stochastic gradient evaluations to achieve a point where the gradient's norm is less than 10^{-10} (10^{-1}).

$0 < \rho < 1$. We can also show that Assumption 1 holds, i.e., $D = \sqrt{2}$, the mapping $F(\mathbf{x}, \mathbf{y}) = (\partial_{\mathbf{x}} f(\mathbf{x}, y), -\partial_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))^\top$ is ρ-weakly monotone and the corresponding MVI problem has a solution $(0, 0, 0, 0)$. In addition, $f(\mathbf{x}, \mathbf{y})$ is L-Lipschitz continuous with modulus 1. For IPP-SGD and IPP-SVRG, we start from $(0.2, 0.1, 0.2, 0.1)$, run both algorithms until reaching a point whose gradient's magnitude is less than a certain threshold ϵ. We consider two different values of ϵ (10^{-10} and 10^{-1}) and report the corresponding number of stochastic gradient evaluations in Table 3. We can see that both algorithms converge. IPP-SVRG converges faster than IPP-SGD when the target accuracy is sufficiently small (e.g., 10^{-10}), and IPP-SGD converges faster than IPP-SGD when the target accuracy is not small enough (e.g. 10^{-1}). This observation is consistent with our theory.

6.2 Training GANs

Next, we report some experimental results to justify the effectiveness of the proposed algorithm for training GANs. Although there are hundreds of variants of GAN formulations, we focus on two popular variants, namely WGAN (Arjovsky et al., 2017) and WGAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017). We conduct the experiments using the same neural network structures for the discriminator and generator as in the original implementations on two datasets (CIFAR10 and LSUN Bedroom). We use two activation functions in the networks, namely ReLU and ELU (Clevert et al., 2015), with the latter being a smooth function which is consistent with our assumption.

We consider two stochastic methods for implementing the subroutine ApproxSVI, namely primal-dual SGD and primal-dual Adam, and refer to our algorithms as IPP-SGD and IPP-Adam. For both of them, the update for the discriminator parameters and for the generator parameters is conducted simultaneously. We compare IPP-SGD with two baselines, namely SGD (5d) and SGD, and compare IPP-Adam with two baselines, namely, Adam (5d) and Adam. 5d means that 5 steps of updates for the discriminator parameters is conducted before 1 step update of the generator parameters, which is considered in their

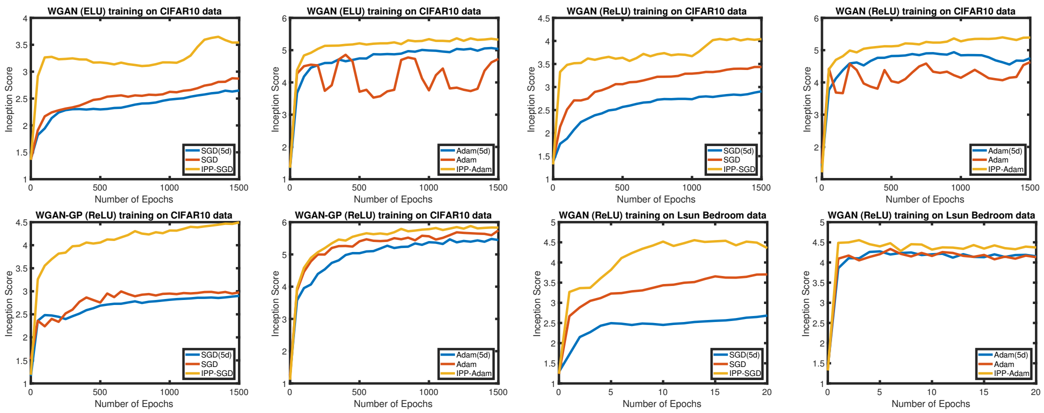

Figure 1: Comparison of different methods for solving WGAN and WGAN-GP.

original papers (Arjovsky et al., 2017; Gulrajani et al., 2017). In SGD and Adam only 1 step of update is performed for the discriminator and the generator simultaneously.

The performance of algorithms is evaluated by using the inception score (Salimans et al., 2016) of the generated images. The step size of SGD and SGD (5d) is set as $\eta_0/\sqrt{t}$ with t being the number of performed updates as in standard theory for convex-concave problems. For IPP-SGD, the step size and the number of inner iterations at k-th stage are set to be $\eta_0/(k+1)$ and $T_0(k+1)^2$ according to our theory, respectively. For IPP-Adam, the step size and the number of iterations at k-th stage are set to be $\eta_0/(k+1)$ and a large value T for simplicity, respectively. All parameters including η_0, T_0, T, γ are tuned for each algorithm separately to obtain the best performance.

CIFAR10 For CIFAR10 data, we tune the initial stepsize of SGD(5d) from $\{10^{-1}, 5\times 10^{-2}, 10^{-2}, 5\times 10^{-3}, 10^{-3}, 5\times 10^{-4}, 10^{-4}\}$, choose the one with best performance, and then use the same initial stepsize for SGD and IPP-SGD. For SGD(5d) and SGD, the stepsize at iteration t is set to be $\eta_0/\sqrt{t}$, where η_0 is initial stepsize. For IPP-SGD, the stepsize and the number of inner iterations at k-th stage are set to be $\eta_0/(k+1)$ and $T_0(k+1)^2$ respectively, where T_0 is tuned from $\{5000 : 5000 : 60000\}$. In our experiment on CIFAR10 data, the tuned initial stepsize is 0.1 for WGAN (ReLU and ELU) and 0.01 for WGAN-GP (ReLU). For Adam(5d), Adam and IPP-Adam, we choose the initial stepsize according to the original implementations for WGAN and WGAN-GP. For IPP-Adam, the stepsize and the number of passes of CIFAR10 data (with batchsize 64) at k-th stage are set to be $\eta_0'/(k+1)$ and c' respectively, where η_0' is the initial stepsize for Adam (5d) and c' is tuned from $\{50 : 50 : 400\}$. Note that the number of iterations for each stage is $T = 50000c'/64$. For our proposed methods (IPP-Adam and IPP-SGD), we tune the weakly monotone parameter from $\{10^{-3}, 10^{-4}\}$.

LSUN Bedroom For LSUN Bedroom dataset, we use the same strategy as we used in CIFAR10 data for SGD(5d), SGD and IPP-SGD, Adam(5d) and Adam. The tuned initial step size for SGD(5d) is 0.01 for WGAN (ReLU and ELU). For IPP-Adam, we use the same initial stepsize as Adam(5d) and tune the number of inner iterations in a different range. Specifically, it is tuned from $\{5000 : 5000 : 50000\}$.

The inception score vs the number of epochs passing the data for solving WGAN and WGAN-GP on the two datasets are plotted in Figure 1. We can see that in all cases, the proposed IPP methods outperform their plain versions used in previous studies. We also notice that using 5 steps of updates for the discriminator before 1 step of update for the generator is observed to be only effective for training WGAN by Adam. This strategy is not that effective for training WGAN-GP or SGD. We also notice that using ADAM in IPP converges faster than using SGD in IPP.

7. Conclusion

In this paper, we have presented the first non-asymptotic convergence result for finding a nearly stationary point of non-convex non-concave saddle-point problems. Our analysis is built on the tool of variational inequalities. An inexact proximal point method is presented with different variations that employ different algorithms for solving the constructed strongly monotone variational inequalities. Synthetic experiments verify our theory and experiments on training two variants of generative adversarial networks demonstrate the effectiveness of the proposed methods.

8. Acknowledgments

We would like to thank anonymous reviewers for their helpful comments. This work was partially supported by NSF #1933212, NSF CAREER Award #1844403.

Appendix A. Proofs of Lemmas

Proof [Proof of Lemma 3] Suppose $F(\mathbf{z}) = (\partial_x f(\mathbf{x},\mathbf{y}), \partial_y[-f(\mathbf{x},\mathbf{y})])$ is ρ-weakly monotone. Let $\mathbf{z}=(\mathbf{x},\mathbf{y})$ and $\mathbf{z}'=(\mathbf{x}',\mathbf{y})$. By definition of weak monotonicity, we have for any $\mathbf{y}\in\mathcal{Y}$, any $\mathbf{x},\mathbf{x}'\in\mathcal{X}$, any $\boldsymbol{\xi}_x\in\partial_x f(\mathbf{x},\mathbf{y})$, any $\boldsymbol{\xi}_x'\in\partial_x f(\mathbf{x}',\mathbf{y})$, any $-\boldsymbol{\xi}_y\in\partial_y[-f(\mathbf{x},\mathbf{y})]$, any $-\boldsymbol{\xi}_y'\in\partial_y[-f(\mathbf{x}',\mathbf{y})]$

$$\left\langle \boldsymbol{\xi}_x-\boldsymbol{\xi}_x', \mathbf{x}-\mathbf{x}'\right\rangle = \left\langle \boldsymbol{\xi}_x-\boldsymbol{\xi}_x', \mathbf{x}-\mathbf{x}'\right\rangle - \left\langle \boldsymbol{\xi}_y-\boldsymbol{\xi}_y', \mathbf{y}-\mathbf{y}\right\rangle = \left\langle \boldsymbol{\xi}-\boldsymbol{\xi}', \mathbf{z}-\mathbf{z}'\right\rangle \geq -\rho\|\mathbf{x}-\mathbf{x}'\|_2$$

where $\boldsymbol{\xi}=(\boldsymbol{\xi}_x,-\boldsymbol{\xi}_y)$ and $\boldsymbol{\xi}'=(\boldsymbol{\xi}_x',-\boldsymbol{\xi}_y')$. This implies that $f(\mathbf{x},\mathbf{y})$ is ρ-weakly convex in $\mathbf{x}$ for any $\mathbf{y}\in\mathcal{Y}$. Similarly, we can show that $f(\mathbf{x},\mathbf{y})$ is ρ-weakly concave in $\mathbf{y}$ for any $\mathbf{x}\in\mathcal{X}$.

Suppose $f(\mathbf{x},\mathbf{y})$ is ρ-weakly-convex-weakly-concave. Now let $\boldsymbol{\xi}_x\in\partial_x f(\mathbf{x},\mathbf{y})$, any $\boldsymbol{\xi}_x'\in\partial_x f(\mathbf{x}',\mathbf{y}')$, any $-\boldsymbol{\xi}_y\in\partial_y[-f(\mathbf{x},\mathbf{y})]$, any $-\boldsymbol{\xi}_y'\in\partial_y[-f(\mathbf{x}',\mathbf{y}')]$. We have

$$\begin{aligned}
\left\langle \boldsymbol{\xi}_x-\boldsymbol{\xi}_x', \mathbf{x}-\mathbf{x}'\right\rangle &\geq f(\mathbf{x},\mathbf{y}) - f(\mathbf{x}',\mathbf{y}) + f(\mathbf{x}',\mathbf{y}') - f(\mathbf{x},\mathbf{y}') - \rho\|\mathbf{x}-\mathbf{x}'\|^2 \\
\left\langle \boldsymbol{\xi}_y'-\boldsymbol{\xi}_y, \mathbf{y}-\mathbf{y}'\right\rangle &\geq -f(\mathbf{x},\mathbf{y}) + f(\mathbf{x},\mathbf{y}') - f(\mathbf{x}',\mathbf{y}') + f(\mathbf{x}',\mathbf{y}) - \rho\|\mathbf{y}-\mathbf{y}'\|^2
\end{aligned}$$

for any $\mathbf{x},\mathbf{x}'\in\mathcal{X}$, any $\mathbf{y},\mathbf{y}\in\mathcal{Y}$, any $\boldsymbol{\xi}_x\in\partial_x f(\mathbf{x},\mathbf{y})$, any $\boldsymbol{\xi}_x'\in\partial_x f(\mathbf{x}',\mathbf{y}')$, any $-\boldsymbol{\xi}_y\in\partial_y[-f(\mathbf{x},\mathbf{y})]$, any $-\boldsymbol{\xi}_y'\in\partial_y[-f(\mathbf{x}',\mathbf{y}')]$. Adding these two inequalities together, we have

$$\left\langle \boldsymbol{\xi}-\boldsymbol{\xi}', \mathbf{z}-\mathbf{z}'\right\rangle \geq -\rho\|\mathbf{x}-\mathbf{x}'\|_2$$

where $\boldsymbol{\xi}=(\boldsymbol{\xi}_x,-\boldsymbol{\xi}_y)$ and $\boldsymbol{\xi}'=(\boldsymbol{\xi}_x',-\boldsymbol{\xi}_y')$. This means $F(\mathbf{z})$ is ρ-weakly monotone. ■

Proof [Proof of Lemma 4] Given any $\mathbf{z}$ and $\mathbf{z}'$ in $\mathcal{Z}$, any $\boldsymbol{\xi}\in F(\mathbf{z})$ and any $\boldsymbol{\xi}'\in F(\mathbf{z}')$, we have

$$\begin{aligned}
&\left\langle \boldsymbol{\xi} + \frac{1}{\gamma}(\mathbf{z}-\mathbf{w}) - \boldsymbol{\xi}' - \frac{1}{\gamma}(\mathbf{z}'-\mathbf{w}), \mathbf{z}-\mathbf{z}'\right\rangle \\
\geq\ & \langle \boldsymbol{\xi}-\boldsymbol{\xi}', \mathbf{z}-\mathbf{z}'\rangle + \frac{1}{\gamma}\|\mathbf{z}-\mathbf{z}'\|^2 \geq \left(\frac{1}{\gamma}-\rho\right)\|\mathbf{z}-\mathbf{z}'\|^2,
\end{aligned}$$

where the second inequality is because of the ρ-weakly monotonicity of F. Since $F_{\mathbf{w}}^{\gamma}(\mathbf{z})$ consists of all vectors like $\boldsymbol{\xi}+\frac{1}{\gamma}(\mathbf{z}-\mathbf{w})$ with $\boldsymbol{\xi}\in F(\mathbf{z})$, we conclude that $F_{\mathbf{w}}^{\gamma}$ is $(\frac{1}{\gamma}-\rho)$-strongly monotone. ■

Proof [Proof of Lemma 6] Since $\bar{\mathbf{w}}$ is the solution to $\text{SVI}(F_{\mathbf{w}}^{\gamma},\mathcal{Z})$ and $F_{\mathbf{w}}$ is strongly monotone, then $\bar{\mathbf{w}}$ is the min-max saddle-point of the convex-concave problem $\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{u}\|^2 - \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{v}\|^2$ (Nemirovski, 2015). Denote by $f_{\mathbf{w}}(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{u}\|^2 - \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{v}\|^2$. Then

$$\mathbf{0}\in\partial_x[f_{\mathbf{w}}(\bar{\mathbf{u}},\bar{\mathbf{v}}) + \mathbf{1}_{\mathcal{X}}(\bar{\mathbf{u}})], \quad \mathbf{0}\in\partial_y[-f_{\mathbf{w}}(\bar{\mathbf{u}},\bar{\mathbf{v}}) + \mathbf{1}_{\mathbf{y}}(\bar{\mathbf{v}})]$$

which means there exist $\bar{\boldsymbol{\xi}}_x\in\partial_x[f(\bar{\mathbf{u}},\bar{\mathbf{v}}) + \mathbf{1}_{\mathcal{X}}(\bar{\mathbf{u}})]$ and $-\bar{\boldsymbol{\xi}}_y\in\partial_y[-f(\bar{\mathbf{u}},\bar{\mathbf{v}}) + \mathbf{1}_{\mathbf{y}}(\bar{\mathbf{v}})]$ such that

$$\|\bar{\boldsymbol{\xi}}\| = \frac{1}{\gamma}\|\bar{\mathbf{w}}-\mathbf{w}\|, \quad \bar{\boldsymbol{\xi}} = (\bar{\boldsymbol{\xi}}_x, -\bar{\boldsymbol{\xi}}_y),$$

which completes the proof. ■

Proof [Proof of Lemma 10] We use the following notations $\mathbf{z}^{(t)} = \widehat{\mathbf{z}}$, $\mathbf{z}^{(t+1)} = \bar{\mathbf{z}} = \text{Proj}_{\mathcal{Z}}(\mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)}))$. For any $\mathbf{z}$, we define

$$\phi_\eta(\mathbf{z}) := F(\mathbf{z}) - F(\mathbf{z}_+) + \frac{1}{\eta}[\mathbf{z}_+ - \mathbf{z}],$$

where $\mathbf{z}_+ = \text{Proj}_{\mathcal{Z}}(\mathbf{z} - \eta F_k(\mathbf{z}))$. Then we have $\|\phi_\eta(\mathbf{z})\| \leq (L + \eta^{-1})\|\mathbf{z} - \mathbf{z}_+\|$ and thus

$$\begin{aligned} \max_{\mathbf{z}\in\mathcal{Z}} F(\mathbf{z}^{(t+1)})^\top(\mathbf{z}^{(t+1)} - \mathbf{z}) &= \max_{\mathbf{z}\in\mathcal{Z}} (F(\mathbf{z}^{(t+1)}) + \phi_\eta(\mathbf{z}^{(t)}) - \phi_\eta(\mathbf{z}^{(t)}))^\top(\mathbf{z}^{(t+1)} - \mathbf{z}) \\ &\leq \max_{\mathbf{z}\in\mathcal{Z}} -\phi_\eta(\mathbf{z}^{(t)})^\top(\mathbf{z}^{(t)} - \mathbf{z}) \leq D(L + \eta^{-1})\|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}\|, \end{aligned} \tag{15}$$

where the first inequality uses the optimality condition of $\mathbf{z}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{z} - (\mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)}))\|^2$ that says

$$\left[(\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)})/\eta + F(\mathbf{z}^{(t)}))\right]^\top (\mathbf{z} - \mathbf{z}^{(t+1)}) = \left[F(\mathbf{z}^{(t+1)}) + \phi_\eta(\mathbf{z}^{(t)})\right]^\top (\mathbf{z} - \mathbf{z}^{(t+1)}) \geq 0, \quad \forall \mathbf{z} \in \mathcal{Z}.$$

To continue, we define $\mathbf{y}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{z} - (\mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t+1)}))\|^2$. According to Lemma 5 in (Dang and Lan, 2015), we have

$$(1 - L^2\eta^2)\frac{\|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}\|^2}{2} \leq \frac{\|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2 - \|\mathbf{y}^{(t+1)} - \mathbf{w}_*\|^2}{2}. \tag{16}$$

Plugging the value of η gives

$$\|\widehat{\mathbf{z}} - \bar{\mathbf{z}}\| \leq \sqrt{2}\|\widehat{\mathbf{z}} - \mathbf{w}_*\|.$$

Combining this inequality with (15) we have

$$\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top(\bar{\mathbf{z}} - \mathbf{z}) \leq DL(1 + \sqrt{2})\sqrt{2}\|\widehat{\mathbf{z}} - \mathbf{w}_*\|$$

■

Appendix B. Weakly-Convex-Weak-Concave Examples

In this section, we present some examples of the min-max problem whose objective function is weakly-convex and weakly-concave.

Example 1: When $f(\mathbf{x}, \mathbf{y})$ is L-smooth function in terms of $\mathbf{x}$ when fixing $\mathbf{y}$ and L-smooth function in terms of $\mathbf{y}$ when fixing $\mathbf{x}$, it is L weakly-convex-weakly-concave. This kind of problems can be found in training GAN Arjovsky et al. (2017) and reinforcement learning Dai et al. (2018). Given the smoothness, we have

$$f(\mathbf{x}, \mathbf{y}) \geq f(\mathbf{x}', \mathbf{y}) + \nabla_x f(\mathbf{x}', \mathbf{y})^\top(\mathbf{x} - \mathbf{x}') - \frac{L}{2}\|\mathbf{x} - \mathbf{x}'\|^2, \forall \mathbf{y} \in \mathcal{Y}$$

which gives us

$$f(\mathbf{x},\mathbf{y}) + \frac{L}{2}\|\mathbf{x}\|^2 \geq f(\mathbf{x}',\mathbf{y}) + \frac{L}{2}\|\mathbf{x}'\|^2 + (\nabla_x f(\mathbf{x}',\mathbf{y}) + L\mathbf{x}')^\top(\mathbf{x}-\mathbf{x}'), \forall \mathbf{y}\in\mathcal{Y},$$

which means $f(\mathbf{x},\mathbf{y}) + \frac{L}{2}\|\mathbf{x}\|^2$ is convex in terms of $\mathbf{x}$ for any fixed $\mathbf{y}\in\mathcal{Y}$. Similarly, we can prove $f(\mathbf{x},\mathbf{y}) - \frac{L}{2}\|\mathbf{y}\|^2$ is a concave function in terms of $\mathbf{y}$ for any $\mathbf{x}\in\mathcal{X}$.

Example 2: Let us consider $f(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top A\mathbf{y} + \phi(g(\mathbf{x})) - \psi(h(\mathbf{y}))$ where $\phi(\cdot)$ and $\psi(\cdot)$ are Lipschitz continuous convex functions, $g(\mathbf{x})$ and $h(\mathbf{y})$ are smooth mappings. Following (Drusvyatskiy and Paquette, 2017), we can prove that $f(\mathbf{x},\mathbf{y})$ is weakly convex in terms of $\mathbf{x}$ when fixing $\mathbf{x}$, and weakly concave in terms of $\mathbf{y}$ when fixing $\mathbf{y}$. More specifically, by the convexity of $\phi(\cdot)$, we have

$$\phi(g(\mathbf{x})) \geq \phi(g(\mathbf{x}')) + \nabla\phi(g(\mathbf{x}'))^\top(g(\mathbf{x}) - g(\mathbf{x}'))$$

By the smoothness of $g(\mathbf{x})$, we know there exists $L>0$ such that

$$\|g(\mathbf{x}) - g(\mathbf{x}') - \nabla g(\mathbf{x}')(\mathbf{x}-\mathbf{x}')\| \leq \frac{L}{2}\|\mathbf{x}-\mathbf{x}'\|^2$$

Combining the above two inequalities we have

$$\phi(g(\mathbf{x})) \geq \phi(g(\mathbf{x}')) + \nabla\phi(g(\mathbf{x}'))^\top\nabla g(\mathbf{x}')(\mathbf{x}-\mathbf{x}') - \frac{L\|\nabla\phi(g(\mathbf{x}'))\|}{2}\|\mathbf{x}-\mathbf{x}'\|^2$$

Since ϕ is Lipschitz continuous, there exists $M>0$ such that $\|\nabla\phi(g(\mathbf{x}'))\| \leq M$. As a result,

$$\phi(g(\mathbf{x})) \geq \phi(g(\mathbf{x}')) + \nabla\phi(g(\mathbf{x}'))^\top\nabla g(\mathbf{x}')(\mathbf{x}-\mathbf{x}') - \frac{LM}{2}\|\mathbf{x}-\mathbf{x}'\|^2.$$

This proves the weak convexity of $\phi(g(\mathbf{x}))$. Similarly we can prove the weak convexity of $\psi(h(\mathbf{y}))$. Since $\mathbf{x}^\top A\mathbf{y}$ is convex in terms of $\mathbf{x}$ when fixing $\mathbf{y}$ and concave in terms of $\mathbf{y}$ when fixing $\mathbf{x}$. Thus, we have weak-convexity and weak-concavity of $f(\mathbf{x},\mathbf{y})$.

It is notable that $f(\mathbf{x},\mathbf{y})$ is not necessarily smooth.

Example 3: Let us consider $f(\mathbf{x},\mathbf{y}) = \phi(g(\mathbf{x}) - h(\mathbf{y}))$ where $\phi(\cdot)$ is a non-decreasing smooth function, $g(\mathbf{x})$ and $h(\mathbf{y})$ are Lipschitz continuous convex functions. Following Xu et al. (2018), it can be proved that $f(\mathbf{x},\mathbf{y})$ is weakly convex in terms of $\mathbf{x}$ when fixing $\mathbf{y}$, and weakly concave in terms of $\mathbf{y}$ when fixing $\mathbf{x}$. The following inequalities hold for any $\mathbf{x},\mathbf{x}'\in\mathcal{X},\mathbf{y},\mathbf{y}'\in\mathcal{Y}$. In fact, by the smoothness of ϕ, there exists $L>0$ such that

$$\phi(g(\mathbf{x}) - h(\mathbf{y})) \geq \phi(g(\mathbf{x}') - h(\mathbf{y})) + \phi'(g(\mathbf{x}') - h(\mathbf{y}))(g(\mathbf{x}) - g(\mathbf{x}')) - \frac{L}{2}|g(\mathbf{x}) - g(\mathbf{x}')|^2,$$

Since $g(\mathbf{x})$ is convex and Lipschitz continuous, we have

$$g(\mathbf{x}) - g(\mathbf{x}') \geq \nabla g(\mathbf{x}')^\top(\mathbf{x}-\mathbf{x}'), \quad |g(\mathbf{x}) - g(\mathbf{x}')| \leq G_x\|\mathbf{x}-\mathbf{x}'\|$$

Noting that ϕ is non-decreasing function with $\phi'(\cdot)\geq 0$, then we have

$$\phi(g(\mathbf{x}) - h(\mathbf{y})) \geq \phi(g(\mathbf{x}') - h(\mathbf{y})) + \phi'(g(\mathbf{x}') - h(\mathbf{y}))\nabla g(\mathbf{x}')^\top(\mathbf{x}-\mathbf{x}') - \frac{LG_x^2}{2}\|\mathbf{x}-\mathbf{x}'\|^2,$$

which implies weak-convexity in terms of $\mathbf{x}$ when fixing $\mathbf{y}$. Similarly,

$$\phi(g(\mathbf{x})-h(\mathbf{y})) \leq \phi(g(\mathbf{x})-h(\mathbf{y}')) - \phi'(g(\mathbf{x})-h(\mathbf{y}'))(h(\mathbf{y})-h(\mathbf{y}')) + \frac{L}{2}|h(\mathbf{y})-h(\mathbf{y}')|^2,$$

Since $h(\mathbf{x})$ is convex and Lipschitz continuous, we have

$$h(\mathbf{y})-h(\mathbf{y}') \geq \nabla h(\mathbf{y}')^\top(\mathbf{y}-\mathbf{y}'), \quad |h(\mathbf{x})-h(\mathbf{x}')| \leq G_y\|\mathbf{y}-\mathbf{y}'\|$$

Then we have

$$\phi(g(\mathbf{x})-h(\mathbf{y})) \leq \phi(g(\mathbf{x})-h(\mathbf{y}')) - \phi'(g(\mathbf{x})-h(\mathbf{y}'))\nabla h(\mathbf{y}')^\top(\mathbf{y}-\mathbf{y}') + \frac{LG_y^2}{2}\|\mathbf{y}-\mathbf{y}'\|^2,$$

which implies weak-concavity in terms of $\mathbf{y}$ when fixing $\mathbf{x}$.

It also is notable that $f(\mathbf{x},\mathbf{y})$ is not necessarily smooth. This problem can be found in robust statistics (Audibert and Catoni, 2011).

Example 4 (An Example Satisfying MVI Condition): We consider $f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{x}^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top A\mathbf{y}$, $\mathcal{X} = \{(x_1,x_2)\in\mathbb{R}^2 : 0\leq x_2\leq x_1\leq 1\}$, $\mathcal{Y} = \{(y_1,y_2)\in\mathbb{R}^2 : 0\leq y_2\leq y_1\leq 1\}$, where $A = \mathrm{diag}(1,-\rho)$, $0<\rho<1$. It is not difficult to show that Assumption 1 holds, i.e., $D=\sqrt{2}$, the mapping $F(\mathbf{x},\mathbf{y}) = (\partial_{\mathbf{x}} f(\mathbf{x},y), -\partial_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))^\top$ is ρ-weakly monotone and the corresponding MVI problem has a solution $(0,0,0,0)$. In addition, $F(\mathbf{x},\mathbf{y})$ is Lipschitz continuous with modulus $L=1$.

Appendix C. Proof of Theorem 7

Let $\mathbb{E}_k$ be the conditional expectation conditioning on all the stochastic events until $\mathbf{z}_k$ is generated. Let $\bar{\mathbf{z}}_k$ be the unique solution of SVI$(F_k,\mathcal{Z})$ where F_k is defined in Algorithm 1. This means

$$\exists\bar{\boldsymbol{\xi}}_k \in F_k(\bar{\mathbf{z}}_k) \text{ s.t. } \bar{\boldsymbol{\xi}}_k^\top(\mathbf{z}-\bar{\mathbf{z}}_k)\geq 0,\ \forall\mathbf{z}\in\mathcal{Z}. \tag{17}$$

By the assumption on ApproxSVI$(F_k,\mathcal{Z},\mathbf{z}_k,\eta_k,T_k)$ and the fact that ApproxSVI only depends on all the previous stochastic events thought $\mathbf{z}_k$, we have

$$\exists\boldsymbol{\xi}_{k+1}\in F_k(\mathbf{z}_{k+1}) \text{ s.t. } \mathbb{E}_k[\boldsymbol{\xi}_{k+1}^\top(\mathbf{z}_{k+1}-\mathbf{z})] \leq \frac{c}{k+1},\ \forall\mathbf{z}\in\mathcal{Z}. \tag{18}$$

By the $(\gamma^{-1}-\rho)$-strong monotonicity of F_k, we have

$$(\gamma^{-1}-\rho)\mathbb{E}_k\|\mathbf{z}_{k+1}-\bar{\mathbf{z}}_k\|^2 \leq \mathbb{E}_k[(\boldsymbol{\xi}_{k+1}-\bar{\boldsymbol{\xi}}_k)^\top(\mathbf{z}_{k+1}-\bar{\mathbf{z}}_k)] \leq \frac{c}{k+1} \tag{19}$$

where the second inequality is obtained using (17) with $\mathbf{z}=\mathbf{z}_{k+1}$ and using (18) with $\mathbf{z}=\bar{\mathbf{z}}_k$.

Let $\mathbf{z}_*$ be a solution to MVI$(F,\mathcal{Z})$, meaning that $\boldsymbol{\xi}^\top(\mathbf{z}-\mathbf{z}_*)\geq 0$ for any $\mathbf{z}\in\mathcal{Z}$ and any $\boldsymbol{\xi}\in F(\mathbf{z})$. Note that such a solution exists by Assumption 1. According to the definition of $F_k(\mathbf{z}_{k+1})$ and the fact that $\boldsymbol{\xi}_{k+1}\in F_k(\mathbf{z}_{k+1})$, we have $\boldsymbol{\xi}_{k+1}-\gamma^{-1}(\mathbf{z}_{k+1}-\mathbf{z}_k)\in F(\mathbf{z}_{k+1})$ so that

$$(\boldsymbol{\xi}_{k+1}-\gamma^{-1}(\mathbf{z}_{k+1}-\mathbf{z}_k))^\top(\mathbf{z}_{k+1}-\mathbf{z}_*)\geq 0$$

by the definition of $\mathbf{z}_*$. This inequality and (18) with $\mathbf{z} = \mathbf{z}_*$ together imply

$$\mathbb{E}_k[(\mathbf{z}_k - \mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}_*)] \geq \gamma\mathbb{E}_k[\boldsymbol{\xi}_{k+1}^\top(\mathbf{z}_* - \mathbf{z}_{k+1})] \geq -\frac{\gamma c}{k+1}.$$

As a result, we have

$$\begin{aligned}
& \mathbb{E}_k\|\mathbf{z}_k - \mathbf{z}_*\|^2 \\
= \ & \mathbb{E}_k\|\mathbf{z}_k - \mathbf{z}_{k+1} - \mathbf{z}_{k+1} - \mathbf{z}_*\|^2 \\
= \ & \mathbb{E}_k\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2 + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2 + 2\mathbb{E}_k[(\mathbf{z}_k - \mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}_*)] \\
\geq \ & \mathbb{E}_k\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2 + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2 - \frac{2\gamma c}{k+1} \\
= \ & \mathbb{E}_k\|\mathbf{z}_k - \bar{\mathbf{z}}_k\|^2 + \mathbb{E}_k\|\mathbf{z}_{k+1} - \bar{\mathbf{z}}_k\|^2 + 2\mathbb{E}_k[(\mathbf{z}_k - \bar{\mathbf{z}}_k)^\top(\bar{\mathbf{z}}_k - \mathbf{z}_{k+1})] + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2 - \frac{2\gamma c}{k+1} \\
\geq \ & \mathbb{E}_k\|\mathbf{z}_k - \bar{\mathbf{z}}_k\|^2 + \mathbb{E}_k\|\mathbf{z}_{k+1} - \bar{\mathbf{z}}_k\|^2 - a\mathbb{E}_k\|\mathbf{z}_k - \bar{\mathbf{z}}_k\|^2 - a^{-1}\mathbb{E}_k\|\bar{\mathbf{z}}_k - \mathbf{z}_{k+1}\|^2 + \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2 - \frac{2\gamma c}{k+1},
\end{aligned}$$

where the last inequality is by Young's inequality and a constant $a \in (0, 1)$. Rearranging the inequality above gives

$$\begin{aligned}
(1-a)\mathbb{E}_k\|\mathbf{z}_k - \bar{\mathbf{z}}_k\|^2 \ \leq \ & (\mathbb{E}_k\|\mathbf{z}_k - \mathbf{z}_*\|^2 - \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2) + (a^{-1} - 1)\mathbb{E}_k\|\bar{\mathbf{z}}_k - \mathbf{z}_{k+1}\|^2 + \frac{2\gamma c}{k+1} \\
\leq \ & (\mathbb{E}_k\|\mathbf{z}_k - \mathbf{z}_*\|^2 - \mathbb{E}_k\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2) + \left(\frac{a^{-1} - 1}{\gamma^{-1} - \rho} + 2\gamma\right)\frac{c}{k+1},
\end{aligned}$$

where the second inequality holds because of (19). Let $\theta_{-1} = 0$. Multiplying both sides of the inequality above by θ_k, taking expectation over all random events, and taking summation over $k = 0, 1, \ldots, K-1$, we have

$$\begin{aligned}
& \sum_{k=0}^{K-1}(1-a)\theta_k\mathbb{E}\|\mathbf{z}_k - \bar{\mathbf{z}}_k\|^2 \\
\leq \ & \sum_{k=0}^{K-1}\theta_k(\mathbb{E}\|\mathbf{z}_k - \mathbf{z}_*\|^2 - \mathbb{E}\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2) + \left(\frac{a^{-1} - 1}{\gamma^{-1} - \rho} + 2\gamma\right)\sum_{k=0}^{K-1}\frac{c\theta_k}{k+1} \\
= \ & \sum_{k=0}^{K-1}(\theta_{k-1}\mathbb{E}\|\mathbf{z}_k - \mathbf{z}_*\|^2 - \theta_k\mathbb{E}\|\mathbf{z}_{k+1} - \mathbf{z}_*\|^2) + \sum_{k=0}^{K-1}(\theta_k - \theta_{k-1})\mathbb{E}\|\mathbf{z}_k - \mathbf{z}_*\|^2 + \left(\frac{a^{-1} - 1}{\gamma^{-1} - \rho} + 2\gamma\right)\sum_{k=0}^{K-1}\frac{c\theta_k}{k+1} \\
= \ & \theta_{-1}\mathbb{E}\|\mathbf{z}_0 - \mathbf{z}_*\|^2 - \theta_{K-1}\mathbb{E}\|\mathbf{z}_K - \mathbf{z}_*\|^2 + (\theta_{K-1} - \theta_{-1})D^2 + \left(\frac{a^{-1} - 1}{\gamma^{-1} - \rho} + 2\gamma\right)\sum_{k=0}^{K-1}\frac{c\theta_k}{k+1} \\
\leq \ & \theta_{K-1}D^2 + \left(\frac{a^{-1} - 1}{\gamma^{-1} - \rho} + 2\gamma\right)\sum_{k=0}^{K-1}\frac{c\theta_k}{k+1}
\end{aligned} \tag{20}$$

Given that $\gamma = 1/(2\rho)$ and the definition of τ, by setting $a = 1/2$ and dividing (20) by $\frac{1}{2}\sum_{k=0}^{K-1}\theta_k$, we have

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \frac{2D^2\theta_{K-1}}{\sum_{k=0}^{K-1}\theta_k} + \frac{4\sum_{k=0}^{K-1}\frac{c\theta_k}{k+1}}{\rho\sum_{k=0}^{K-1}\theta_k}. \tag{21}$$

Since $\alpha > 0$, standard calculus yields

$$\sum_{k=1}^{K} k^{\alpha} \geq \int_{0}^{K} x^{\alpha} dx = \frac{1}{\alpha+1} K^{\alpha+1}$$

$$\sum_{k=1}^{K} k^{\alpha-1} \leq K K^{\alpha-1} = K^{\alpha}, \quad \text{if } \alpha \geq 1$$

$$\sum_{k=1}^{K} k^{\alpha-1} \leq \int_{0}^{K} x^{\alpha-1} dx = \frac{K^{\alpha}}{\alpha}, \quad \text{if } 0 < \alpha < 1.$$

Applying the fact that $\theta_k = (k+1)^{\alpha}$ and the above inequalities into (21), we have

$$\mathbb{E}[\|\mathbf{z}_{\tau} - \bar{\mathbf{z}}_{\tau}\|^2] \leq \frac{2D^2(\alpha+1)}{K} + \frac{4c(\alpha+1)}{K\rho\alpha^{\mathbf{1}_{\alpha<1}}}.$$

Appendix D. Proof of Proposition 8

The proof is following the standard analysis of stochastic subgradient method. By the updating steps in Algorithm 2, for any $\mathbf{z} \in \mathcal{Z}$, we have

$$\boldsymbol{\zeta}(\mathbf{z}^{(t)})^{\top}(\mathbf{z}^{(t)} - \mathbf{z}) \leq \frac{\|\mathbf{z}^{(t)} - \mathbf{z}\|^2 - \|\mathbf{z}^{(t+1)} - \mathbf{z}\|^2}{2\eta} + \frac{\eta\|\boldsymbol{\zeta}(\mathbf{z}^{(t)}\|^2}{2}$$

Adding the inequalities for $t = 0, 1, 2, ..., T-1$ leads to

$$\sum_{t=0}^{T-1} \boldsymbol{\zeta}(\mathbf{z}^{(t)})^{\top}(\mathbf{z}^{(t)} - \mathbf{z}) \leq \frac{1}{2\eta}||\mathbf{z}^{(0)} - \mathbf{z}||^2 + \frac{\eta}{2}\sum_{t=0}^{T-1} ||\boldsymbol{\zeta}(\mathbf{z}^{(t)})||^2$$

Let τ be a uniform random index from $\{0, 1, 2, ..., T-1\}$. Using the previous inequality, we can show that

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{\xi}^{(\tau)})^{\top}(\mathbf{z}^{(\tau)} - \mathbf{z})] &\leq \frac{||\mathbf{z}^{(0)} - \mathbf{z}||^2}{2\eta T} + \frac{\eta}{2}\mathbb{E}[||\boldsymbol{\zeta}(\mathbf{z}^{(\tau)})||^2] \\
&\leq \frac{D^2}{2\eta T} + \frac{\eta G_k^2}{2}.
\end{aligned}$$

Appendix E. Proof of Proposition 11: Part I for GD

We first show that $\|\mathbf{z}^{(t)} - \mathbf{w}_*\|$ converges to zero linearly where $\mathbf{w}_*$ is a solution to SVI$(F_k, \mathcal{Z})$. The proof is standard and can be found in Nesterov and Scrimali (2011). We include it here only for the completeness. The definition of $\mathbf{w}_*$ implies $\mathbf{w}_* = \text{Proj}_{\mathcal{Z}}(\mathbf{w}_* - \eta F_k(\mathbf{w}_*))$. Using the fact that $\eta = \mu/(2L^2)$ and the non-expansion property of $\text{Proj}_{\mathcal{Z}}(\cdot)$, we have

$$\begin{aligned}
\|\mathbf{z}^{(t+1)} - \mathbf{w}_*\|^2 &\leq \|\mathbf{z}^{(t)} - \eta F_k(\mathbf{z}^{(t)}) - \mathbf{w}_* + \eta F_k(\mathbf{w}_*)\|^2 \\
&= \|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2 - 2\eta(F_k(\mathbf{z}^{(t)}) - F_k(\mathbf{w}_*))^{\top}(\mathbf{z}^{(t)} - \mathbf{w}_*) + \eta^2\|F_k(\mathbf{z}^{(t)}) - F_k(\mathbf{w}_*)\|^2 \\
&\leq (1 - 2\eta\mu + \eta^2 L^2)\|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2 = (1 - 3\mu^2/(4L^2))\|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2 \\
&\leq \exp(-\frac{3}{4\beta^2})\|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2
\end{aligned}$$

where the second inequality holds because of the μ-strong monotonicity and L-Lipschitz continuity of F_k. Applying this inequality for $t = 0, 1, \ldots,$ gives

$$\|\mathbf{z}^{(t+1)} - \mathbf{w}_*\|^2 \le \exp(-\frac{3(t+1)}{4\beta^2})\|\mathbf{z}^{(0)} - \mathbf{w}_*\|^2. \tag{22}$$

Next, we prove that $\max_{\mathbf{z}\in\mathcal{Z}} F_k(\mathbf{z}^{(t+1)})^\top(\mathbf{z}^{(t+1)} - \mathbf{z}) \le \|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2$. The main idea of this proof is originally from Dang and Lan (2015). We need to introduce

$$\phi_\eta(\mathbf{z}) := F_k(\mathbf{z}) - F_k(\mathbf{z}_+) + \frac{1}{\eta}[\mathbf{z}_+ - \mathbf{z}],$$

where $\mathbf{z}_+ = \text{Proj}_{\mathcal{Z}}(\mathbf{z} - \eta F_k(\mathbf{z}))$. Then we have $\|\phi_\eta(\mathbf{z})\| \le (L + \eta^{-1})\|\mathbf{z} - \mathbf{z}_+\|$ and thus

$$\begin{aligned}\max_{\mathbf{z}\in\mathcal{Z}} F_k(\mathbf{z}^{(t+1)})^\top(\mathbf{z}^{(t+1)} - \mathbf{z}) &= \max_{\mathbf{z}\in\mathcal{Z}}(F_k(\mathbf{z}^{(t+1)}) + \phi_\eta(\mathbf{z}^{(t)}) - \phi_\eta(\mathbf{z}^{(t)}))^\top(\mathbf{z}^{(t+1)} - \mathbf{z})\\ &\le \max_{\mathbf{z}\in\mathcal{Z}} -\phi_\eta(\mathbf{z}^{(t)})^\top(\mathbf{z}^{(t)} - \mathbf{z}) \le D(L + \eta^{-1})\|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}\|, \end{aligned} \tag{23}$$

where the first inequality uses the optimality condition of $\mathbf{z}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{z} - (\mathbf{z}^{(t)} - \eta F_k(\mathbf{z}^{(t)}))\|^2$ that says

$$\left[(\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)})/\eta + F_k(\mathbf{z}^{(t)}))\right]^\top (\mathbf{z}-\mathbf{z}^{(t+1)}) = \left[F_k(\mathbf{z}^{(t+1)}) + \phi_\eta(\mathbf{z}^{(t)})\right]^\top (\mathbf{z}-\mathbf{z}^{(t+1)}) \ge 0, \quad \forall \mathbf{z} \in \mathcal{Z}.$$

To continue, we define $\mathbf{y}^{(t+1)} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{z} - (\mathbf{z}^{(t)} - \eta F_k(\mathbf{z}^{(t+1)}))\|^2$. According to Lemma 5 in Dang and Lan (2015), we have

$$(1 - L^2\eta^2)\frac{\|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}\|^2}{2} \le \frac{\|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2 - \|\mathbf{y}^{(t+1)} - \mathbf{w}_*\|^2}{2}. \tag{24}$$

Plugging the value of η into (24) gives

$$\|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}\| \le \sqrt{\frac{1}{(1 - 0.25\beta^{-2})}}\|\mathbf{z}^{(t)} - \mathbf{w}_*\| \tag{25}$$

Finally, using (22), (23) and (25) together, we can show that

$$\begin{aligned}\max_{\mathbf{z}\in\mathcal{Z}} F_k(\mathbf{z}^{(t+1)})^\top(\mathbf{z}^{(t+1)} - \mathbf{z}) &\le DL(1+2\beta)\sqrt{\frac{1}{(1 - 0.25\beta^{-2})}}\|\mathbf{z}^{(t)} - \mathbf{w}_*\|\\ &\le DL(1+2\beta)\sqrt{\frac{\beta^2}{\beta^2 - 0.25}}\exp(-\frac{3t}{8\beta^2})\|\mathbf{z}^{(0)} - \mathbf{w}_*\|\\ &\le DL(1+2\beta)\sqrt{\frac{\beta^2}{\beta^2 - 0.25\beta^2}}\exp(-\frac{3t}{8\beta^2})\|\mathbf{z}^{(0)} - \mathbf{w}_*\|\\ &\le 4D^2L\beta\exp(-\frac{t}{4\beta^2}),\end{aligned}$$

where we use the fact $\beta \ge 1$ in the third and the fourth inequalities. The first conclusion is then proved by setting $t = T - 1$.

Since F_k is μ-strongly monotone, we have

$$\mu||\mathbf{z}^{(T)} - \mathbf{w}_*||^2 \le (F_k(\mathbf{z}^{(T)}) - F_k(\mathbf{w}_*))^\top(\mathbf{z}^{(T)} - \mathbf{w}_*) \le (F(\mathbf{z}^{(T)}))^\top(\mathbf{z}^{(T)} - \mathbf{w}_*)$$

where $\mathbf{w}_*$ denotes a solution to $\text{SVI}(F_k, \mathcal{Z})$, which yields the second conclusion.

Appendix F. Proof of Proposition 11: Part II for EG

It is worth mentioning that the linear convergence of the extragradient for strongly monotone VI in terms of distance to the optimal solution is also proved in (Gidel et al., 2018). For completeness, we present a proof here. We can use the following lemma to prove the linear convergence.

Lemma 16 (Lemma 3.1 Nemirovski (2005)) *Let $\omega(\mathbf{z})$ be a α-strongly convex function with respect to the norm $\|\cdot\|$, whose dual norm is denoted by $\|\cdot\|_*$, and $D(\mathbf{x},\mathbf{z}) = \omega(\mathbf{x}) - (\omega(\mathbf{z}) + (\mathbf{x}-\mathbf{z})^\top \omega'(\mathbf{z}))$ be the Bregman distance induced by function $\omega(\mathbf{x})$. Let Z be a convex compact set, and $U \subseteq Z$ be convex and closed. Let $\mathbf{z} \in Z$, $\gamma > 0$, Consider the points,*

$$\mathbf{x} = \arg\min_{\mathbf{u}\in U} \gamma \mathbf{u}^\top \xi + D(\mathbf{u},\mathbf{z}), \tag{26}$$

$$\mathbf{z}_+ = \arg\min_{\mathbf{u}\in U} \gamma \mathbf{u}^\top \zeta + D(\mathbf{u},\mathbf{z}), \tag{27}$$

then for any $\mathbf{u} \in U$, we have

$$\gamma \zeta^\top (\mathbf{x}-\mathbf{u}) \le D(\mathbf{u},\mathbf{z}) - D(\mathbf{u},\mathbf{z}_+) + \frac{\gamma^2}{\alpha}\|\xi - \zeta\|_*^2 - \frac{\alpha}{2}[\|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{x}-\mathbf{z}_+\|^2]. \tag{28}$$

Proof Following the Lemma 16, we can easily have the following equality for the extragrdient method

$$\begin{aligned}2\eta F(\mathbf{z}^{(t)})^\top (\mathbf{z}^{(t)} - \mathbf{w}_*) \le \|\mathbf{w}_* - \mathbf{w}^{(t)}\|^2 - \|\mathbf{w}_* - \mathbf{w}^{(t+1)}\|^2 + 2\eta^2 \|F(\mathbf{z}^{(t)} - F(\mathbf{w}^{(t)})\|^2 \\ - [\|\mathbf{z} - \mathbf{w}^{(t+1)}\|^2 + \|\mathbf{z} - \mathbf{w}^{(t)}\|^2].\end{aligned}$$

By the strong monotonicity, we have

$$\begin{aligned}\frac{1}{2}\mu\|\mathbf{w}^{(t)} - \mathbf{w}_*\|^2 - \mu\|\mathbf{z}^{(t)} - \mathbf{w}^{(t)}\|^2 \le \mu\|\mathbf{z}^{(t)} - \mathbf{w}_*\|^2 \le (F(\mathbf{z}^{(t)}) - F(\mathbf{w}_*))^\top (\mathbf{z}^{(t)} - \mathbf{w}_*) \\ \le F(\mathbf{z}^{(t)})^\top (\mathbf{z}^{(t)} - \mathbf{w}_*).\end{aligned}$$

Combining the above inequalities, we have

$$\begin{aligned}\eta\mu\|\mathbf{w}^{(t)} - \mathbf{w}_*\|^2 - 2\eta\mu\|\mathbf{z}^{(t)} - \mathbf{w}^{(t)}\|^2 \le \|\mathbf{w}_* - \mathbf{w}^{(t)}\|^2 - \|\mathbf{w}_* - \mathbf{w}^{(t+1)}\|^2 + 2\eta^2\|F(\mathbf{z}^{(t)} - F(\mathbf{w}^{(t)})\|^2 \\ - [\|\mathbf{z} - \mathbf{w}^{(t+1)}\|^2 + \|\mathbf{z} - \mathbf{w}^{(t)}\|^2].\end{aligned}$$

Reorganizing the terms we have

$$\begin{aligned}\|\mathbf{w}_* - \mathbf{w}^{(t+1)}\|^2 &\le (1-\eta\mu)\|\mathbf{w}^{(t)} - \mathbf{w}_*\|^2 + 2\eta\mu\|\mathbf{z}^{(t)} - \mathbf{w}^{(t)}\|^2 + 2\eta^2\|F(\mathbf{z}^{(t)} - F(\mathbf{w}^{(t)})\|^2 \\ &\quad - [\|\mathbf{z} - \mathbf{w}^{(t+1)}\|^2 + \|\mathbf{z} - \mathbf{w}^{(t)}\|^2] \\ &\le (1-\eta\mu)\|\mathbf{w}^{(t)} - \mathbf{w}_*\|^2 + (2\eta^2 L^2 + 2\eta\mu)\|\mathbf{z}^{(t)} - \mathbf{w}^{(t)}\|^2 - \|\mathbf{z}^{(t)} - \mathbf{w}^{(t)}\|^2.\end{aligned}$$

By setting $\eta = \frac{1}{4L}$, we have $2\eta^2 L^2 + 2\eta\mu \le 1$, and then

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}_*\|^2 \le (1-\eta\mu)\|\mathbf{w}^{(t)} - \mathbf{w}_*\|^2 \le \left(1 - \frac{\mu}{4L}\right)^{t+1} \|\mathbf{w}^{(0)} - \mathbf{w}_*\|^2.$$

The conclusion follows by applying Lemma 10.

■

Algorithm 6 Nesterov's Accelerated Method for SVI(F, Z): NA$(F, \mathcal{Z}, \mathbf{w}^{(0)}, T)$

Input: $\mu-$Strongly Monotone and L-Lipschitz continuous Mapping F, set $\mathcal{Z}$, $\mathbf{w}^{(0)} \in \mathcal{Z}$, and an integer $T \geq 1$.
Initialize $\lambda_0 = 1$ and $S_0 = 1$, $\beta = L/\mu$, $\eta = 1/(\sqrt{2}L)$
for $t = 0, ..., T-1$ **do**
  $\mathbf{w}^{(t)} = \arg\max_{\mathbf{x}\in\mathcal{Z}} \sum_{k=0}^{t} \lambda_k (F(\mathbf{z}^{(k)})^\top (\mathbf{z}^{(k)} - \mathbf{x}) - \frac{\mu}{2}\|\mathbf{x} - \mathbf{z}^{(k)}\|^2)$
  $\mathbf{z}^{(t+1)} = \arg\max_{\mathbf{x}\in\mathcal{Z}} F(\mathbf{w}^{(k)})^\top (\mathbf{w}^{(t)} - \mathbf{x}) - \frac{L}{2}\|\mathbf{x} - \mathbf{w}^{(t)}\|^2$
  $\lambda_{t+1} = \frac{S_t}{\beta}$
  $S_{t+1} = S_t + \lambda_{t+1}$
end for
Compute $\widehat{\mathbf{z}}_T = \frac{1}{S_T}\sum_{t=0}^{T} \lambda_t \mathbf{z}^{(t)}$
Output: $\bar{\mathbf{z}} = \mathrm{Proj}_{\mathcal{Z}}\left(\widehat{\mathbf{z}}_T - \eta F(\widehat{\mathbf{z}}_T)\right)$

Appendix G. Other Methods for Solving Strongly Monotone SVI with a Lipschitz Continuous Mapping

G.1 Using the Nesterov's Accelerated Method

We first present the Nesterov's accelerated method Nesterov and Scrimali (2011) for solving strongly monotone SVI in Algorithm 6 and show that it could achieve smaller complexity than the GD method when the condition number L/ρ is large.

Proposition 17 *When $F(\mathbf{z})$ is single-valued and L-Lipschitz continuous. and is μ-strongly monotone, Algorithm 6 guarantees that for any $\mathbf{z} \in \mathcal{Z}$*

$$\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top (\bar{\mathbf{z}} - \mathbf{z}) \leq 6\sqrt{M} DL^2/\mu^{3/2} \exp(-T/2(\beta + 1)) \tag{29}$$

where $\beta = L/\mu$. In addition, we have

$$\mu\|\bar{\mathbf{z}} - \mathbf{w}_*\|^2 \leq 6\sqrt{M} DL^2/\mu^{3/2} \exp(-T/2(\beta + 1))$$

where $\mathbf{w}_*$ denotes a solution to SVI$(F, \mathcal{Z})$, and $M = \max_{\mathbf{z},\mathbf{w}\in\mathcal{Z}} F(\mathbf{w})^\top (\mathbf{z} - \mathbf{w}) + \frac{\mu}{2}\|\mathbf{z} - \mathbf{w}\|^2$.

Proof The proof is following. First, according to Theorem 3 in Nesterov and Scrimali (2011), we have

$$\|\widehat{\mathbf{z}}_T - \mathbf{w}_*\|^2 \leq 2M\beta^2/\mu \exp(-T/(\beta + 1))$$

Following Lemma 10, we have

$$\begin{aligned}\max_{\mathbf{z}\in\mathcal{Z}} F(\bar{\mathbf{z}})^\top (\bar{\mathbf{z}} - \mathbf{z}) &\leq DL(2 + \sqrt{2})\|\widehat{\mathbf{z}}_T - \mathbf{w}_*\| \\ &\leq 6\sqrt{M} DL^2/\mu^{3/2} \exp(-T/2(\beta + 1))\end{aligned}$$

■

Corollary 18 *Suppose Assumption 1 holds and $F(\mathbf{z})$ is single-valued and L-Lipschitz continuous, and Algorithm 6 is used to implement ApproxSVI. Algorithm 1 with $\gamma = 1/(2\rho)$, $\theta_k = (k+1)^\alpha$ with $\alpha > 1$, $T_k = 2(L/\rho + 1)\log(6(k+1)L^2M^{1/2}/(D\rho^{5/2}))$, and a total of stages $K = 6D^2(\alpha+1)/\epsilon^2$ guarantees*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \epsilon^2, \tag{30}$$

and

$$\mathbb{E}[dist^2(0, \partial(f(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau) + 1_{\mathcal{Z}}(\bar{\mathbf{u}}_\tau, \bar{\mathbf{v}}_\tau))] \leq \epsilon^2/\gamma^2 \tag{31}$$

where $\bar{\mathbf{z}}_\tau$ is the solution to SVI($F^\gamma_{\mathbf{z}_\tau}, \mathcal{Z}$). The total iteration complexity is $\widetilde{O}(D^2L/(\rho\epsilon^2))$.

Remark: The total iteration complexity for finding a nearly ϵ-stationary solution of the corresponding min-max saddle-point problem such that $\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq (\gamma\epsilon)^2$ is then $\widetilde{O}(L\rho/\epsilon^2)$. By comparing to the result in Corollary 12, the above result of using Nesterov's accelerated method for solving the strongly monotone subproblems is better by a factor of L/ρ.

Appendix H. Solving Weakly Monotone SVI

As we mentioned in the introduction, the proposed algorithms can be used for solving a more general SVI problem. We present and discuss the results in this section. Recall the SVI problem defined in (3) and the MVI problem defined in (4), where F is not necessarily pertained to any $f(\mathbf{x}, \mathbf{y})$. In the literature of VI Nemirovski (2015), a solution $\mathbf{z}^*$ that satisfies (3) is also called strong solution, and a solution $\mathbf{z}_*$ that satisfies (4) is called a weak solution. When F is monotone, finding a solution for SVI($F, \mathcal{Z}$) is typically a tractable problem Nemirovski (2015). In order to make a SVI problem with a non-monotone set-valued mapping F tractable, we impose the same assumptions as in Assumption 1. Similar (or stronger) assumptions have been used in previous studies for solving non-monotone SVI.

When applying an iterative numerical algorithm to solve SVI($F, \mathcal{Z}$), it is generally hard to guarantee an exact solution for SVI($F, \mathcal{Z}$) after a finite number of iterations. Therefore, an alternative goal is to find an *ϵ-gap solution* for SVI($F, \mathcal{Z}$), namely, a solution $\bar{\mathbf{z}}$ such that

$$\exists \bar{\boldsymbol{\xi}} \in F(\bar{\mathbf{z}}) \text{ s.t. } \max_{\mathbf{z}\in\mathcal{Z}} \langle \bar{\boldsymbol{\xi}}, \bar{\mathbf{z}} - \mathbf{z}\rangle \leq \epsilon. \tag{32}$$

However, without additional assumption on F, finding an *ϵ-gap solution* for SVI($F, \mathcal{Z}$) in finite iterations can be also challenging even if F is monotone. For example, consider the SVI problem of finding $z^* \in [-1, 1]$ such that $\langle \xi^*, z - z^*\rangle \geq 0$ for some $\xi^* \in \partial|z^*|$ and all $z \in [-1, 1]$, which is associated to the convex minimization $\min_{z\in[-1,1]} |z|$ and has a solution at 0. Hence, if $\bar{z}$ is very close to 0 but not 0, we always have $\langle \bar{\xi}, \bar{z} - z\rangle \geq 1$ and $|\bar{\xi}| = 1$ for any $\bar{\xi} \in \partial|\bar{z}|$ and $z = -\text{sign}(\bar{z})$. To address this issue, we introduce the notion of nearly (ϵ, δ)-gap solution to SVI($F, \mathcal{Z}$).

Definition 19 *A point $\mathbf{w} \in \mathcal{Z}$ is called a nearly (ϵ, δ)-gap solution to SVI($F, \mathcal{Z}$) for $\epsilon > 0$ and $\delta > 0$ if there exists $\bar{\mathbf{w}} \in \mathcal{Z}$ and such that*

$$\|\mathbf{w} - \bar{\mathbf{w}}\| \leq \delta, \quad \exists \bar{\boldsymbol{\xi}} \in F(\bar{\mathbf{w}}) \ s.t. \ \max_{\mathbf{z}\in\mathcal{Z}} \langle \bar{\boldsymbol{\xi}}, \bar{\mathbf{w}} - \mathbf{z}\rangle \leq \epsilon.$$

If $\delta = O(\epsilon)$, we simply call this solution as nearly ϵ-gap solution. The lemma below implies that the proposed algorithms can find a nearly ϵ-gap solution for a SVI problem.

Lemma 20 *Let $F_{\mathbf{w}}^{\gamma}$ be defined in (6) for $0 < \gamma < \rho^{-1}$ and $\mathbf{w} \in \mathcal{Z}$ and $\bar{\mathbf{w}}$ be the solution to SVI($F_{\mathbf{w}}^{\gamma}, \mathcal{Z}$). We have*

$$\exists \bar{\boldsymbol{\xi}} \in F(\bar{\mathbf{w}}) \;\; s.t. \;\; \max_{\mathbf{z}\in\mathcal{Z}} \langle \bar{\boldsymbol{\xi}}, \bar{\mathbf{w}} - \mathbf{z} \rangle \leq \frac{D}{\gamma} \|\mathbf{w} - \bar{\mathbf{w}}\|. \tag{33}$$

According to this lemma, if we can find a solution $\mathbf{w} \in \mathcal{Z}$ such that $\|\mathbf{w} - \bar{\mathbf{w}}\| \leq \frac{\gamma\epsilon}{D}$, we will have $\max_{\mathbf{z}\in\mathcal{Z}} \langle \bar{\boldsymbol{\xi}}, \bar{\mathbf{w}} - \mathbf{z} \rangle \leq \epsilon$, namely, $\mathbf{w}$ is $(\gamma\epsilon/D)$-closed to an ϵ-gap solution of SVI($F, \mathcal{Z}$).

Next we present a proposition to show that, when $F(\mathbf{z})$ is single-valued and Lipschitz continuous, a nearly (ϵ, δ)-gap solution with $\delta = O(\epsilon)$ is an $O(\epsilon)$-gap solution to SVI($F, \mathcal{Z}$), and an ϵ-gap solution $\mathbf{w}$ is $O(\sqrt{\epsilon})$-close to the solution of SVI($F_{\mathbf{w}}^{\gamma}, \mathcal{Z}$) for $\gamma \in (0, L^{-1})$.

Proposition 21 *When $F(\mathbf{z})$ is single-valued and L-Lipschitz continuous, the following statements hold:*

- *If $\mathbf{w}$ is a nearly $(\epsilon, c\epsilon)$-gap solution for $c > 0$, then $\mathbf{w}$ is an $((1 + LcD + Mc)\,\epsilon)$-gap solution to SVI($F, \mathcal{Z}$), where $M \equiv \max_{\mathbf{z}\in\mathcal{Z}} \|F(\mathbf{z})\|$.*
- *If $\mathbf{w}$ is an ϵ-gap solution to SVI($F, \mathcal{Z}$), then $\|\mathbf{w} - \widehat{\mathbf{w}}\| \leq \sqrt{\frac{\epsilon}{\gamma^{-1} - L}}$, where $\widehat{\mathbf{w}}$ is the unique solution of SVI($F_{\mathbf{w}}^{\gamma}, \mathcal{Z}$) for $0 < \gamma < L$.*

This proposition indicates that a nearly ϵ-gap solution is the right target when solving SVI($F, \mathcal{Z}$) no matter F is single-valued Lipschitz continuous or set-valued non-Lipschitz.

Now consider the second statement. Suppose $\bar{\mathbf{z}} \in \mathcal{Z}$ is an ϵ-gap solution of SVI($F, \mathcal{Z}$) such that $\max_{\mathbf{z}\in\mathcal{Z}} \langle F(\bar{\mathbf{z}}), \bar{\mathbf{z}} - \mathbf{z} \rangle \leq \epsilon$. Note that $F_{\bar{\mathbf{z}}}^{\gamma}$ is $\left(\frac{1}{\gamma} - L\right)$-strongly monotone for $\gamma \in (0, L^{-1})$ because F is L-Lipschitz continuous. Let $\widehat{\mathbf{z}}$ be the unique solution of SVI($F_{\bar{\mathbf{z}}}^{\gamma}, \mathcal{Z}$) for $\gamma \in (0, L^{-1})$. We then have

$$\left\langle F(\widehat{\mathbf{z}}) + \frac{1}{\gamma}(\widehat{\mathbf{z}} - \bar{\mathbf{z}}), \widehat{\mathbf{z}} - \bar{\mathbf{z}} \right\rangle \leq 0$$

which, by the Lipschitz continuity of F, implies

$$\frac{1}{\gamma}\|\widehat{\mathbf{z}} - \bar{\mathbf{z}}\|^2 \leq \langle F(\widehat{\mathbf{z}}), \bar{\mathbf{z}} - \widehat{\mathbf{z}} \rangle \leq \langle F(\bar{\mathbf{z}}), \bar{\mathbf{z}} - \widehat{\mathbf{z}} \rangle + L\|\bar{\mathbf{z}} - \widehat{\mathbf{z}}\|^2 \leq \epsilon + L\|\bar{\mathbf{z}} - \widehat{\mathbf{z}}\|^2.$$

By reorganizing terms, we have $\left(\frac{1}{\gamma} - L\right)\|\bar{\mathbf{z}} - \widehat{\mathbf{z}}\|^2 \leq \epsilon$ which leads to the conclusion of the second statement.

The next two corollaries summarize the convergence results of Algorithm 1 for the solving the SVI problem (3).

Corollary 22 *For the SVI problem (3), assume that Assumption 1 holds and for any $\mathbf{z} \in \mathcal{Z}$ there exists $\boldsymbol{\zeta}(\mathbf{z}) \in F(\mathbf{z})$ such that $\mathbb{E}[\boldsymbol{\zeta}(\mathbf{z})] \in F(\mathbf{z})$, and $\mathbb{E}\|\boldsymbol{\zeta}(\mathbf{z})\|^2 \leq G^2$. Under the same conditions as in Corollary 9 except for $K = \frac{(16\rho^2D^2 + 4\rho DG)D^2(\alpha+1)}{\epsilon^2}$ in Algorithm 1, we have*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \frac{\gamma^2\epsilon^2}{D^2},$$

$$\exists \boldsymbol{\xi} \in F(\bar{\mathbf{z}}_\tau) \;\; s.t. \;\; \mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} \langle \boldsymbol{\xi}, \bar{\mathbf{z}}_\tau - \mathbf{z} \rangle] \leq \epsilon,$$

where $\bar{\mathbf{z}}_\tau$ is the solution to SVI($F^{\gamma}_{\mathbf{z}_\tau}, \mathcal{Z}$). The total iteration complexity of $O(\frac{1}{\epsilon^6})$.

Remark: The above result establishes the convergence result for finding a nearly $(\epsilon, \frac{\gamma\epsilon}{D})$-gap solution for SVI($F, \mathcal{Z}$). The total iteration complexity can be easily derived as $\sum_{k=1}^K k^2 = O(1/\epsilon^6)$. To our knowledge, this is the first non-asymptotic convergence of stochastic algorithms for solving SVI without the monotone and Lipschitz conditions.

Corollary 23 *For the SVI problem (3), assume that Assumption 1 holds and F is single-valued and L_k-Lipschitz continuous. Under the same conditions as in Corollary 12 except for $K = \frac{16\rho^2 D^4(\alpha+1)}{\epsilon^2}$ in Algorithm 1, we have*

$$\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|^2] \leq \frac{\gamma^2\epsilon^2}{D^2}, \quad \mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} \langle F(\bar{\mathbf{z}}_\tau), \bar{\mathbf{z}}_\tau - \mathbf{z}\rangle] \leq \epsilon$$

where $\bar{\mathbf{z}}_\tau$ is the solution to SVI($F^{\gamma}_{\mathbf{z}_\tau}, \mathcal{Z}$). The total iteration complexity of $O(\log(\frac{1}{\epsilon})\frac{D^4L^2}{\epsilon^2})$.

Remark: One can similarly derive the results based on the Nesterov's accelerated method Nesterov and Scrimali (2011) and the extragradient method Korpelevich (1976), which can improve the above complexity when the condition number $L/\rho \gg 1$ is large. Next, we compare the above result with Dang and Lan (2015) for solving a SVI with L-Lipschitz continuous single-valued mapping, which presents an extragradient algorithm and needs to compute two gradient updates at each iteration and has an iteration complexity of $O(L^2D^4/\epsilon^2)$ for finding an ϵ-gap solution to the SVI. Without additional assumption on the value of ρ, we can set $\rho = L$. Then the iteration complexity of Algorithm 1 for ensuring $\mathbb{E}[\|\mathbf{z}_\tau - \bar{\mathbf{z}}_\tau\|] \leq \epsilon/(LD)$ and $\mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} \langle F(\bar{\mathbf{z}}_\tau), \bar{\mathbf{z}}_\tau - \mathbf{z}\rangle] \leq 2\epsilon$ is $O(L^2D^4/\epsilon^2\log(1/\epsilon))$, which implies that $\mathbb{E}[\max_{\mathbf{z}\in\mathcal{Z}} \langle F(\mathbf{z}_\tau), \mathbf{z}_\tau - \mathbf{z}\rangle] \leq O(\epsilon)$ according to Proposition 21. As we can see our complexity is worse by a $\log(1/\epsilon)$ factor but only needs to perform one gradient update at each iteration.

Discussion: ϵ-gap vs ϵ-stationary. Recall the definition of an ϵ-stationary solution:

$$\text{dist}(\mathbf{0}, \partial(f(\mathbf{x}, \mathbf{y}) + 1_{\mathcal{Z}}(\mathbf{x}, \mathbf{y})) \leq \epsilon, \tag{34}$$

It is an interesting question whether one can derive an ϵ-stationary result for the min-max problem (i.e., (34)) from an ϵ-gap solution that satisfies (32) of the corresponding SVI. We will show that (34) is a stronger result than (32), i.e., ϵ-stationary solution is also an $O(\epsilon)$-gap solution of the corresponding SVI but not vice versa. Let $F(\mathbf{z}) = (\partial_x f(\mathbf{x}, \mathbf{y}), -\partial_y f(\mathbf{x}, \mathbf{y}))^\top$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Suppose a solution $\bar{\mathbf{z}}$ is found such that $\text{dist}^2(0, F(\bar{\mathbf{z}}) + 1_{\mathcal{Z}}(\bar{\mathbf{z}})) \leq \epsilon^2$. Hence, there exists $\bar{\boldsymbol{\xi}} \in F(\bar{\mathbf{z}})$ and $\bar{\boldsymbol{\zeta}} \in \partial 1_{\mathcal{Z}}(\bar{\mathbf{z}})$ such that $\|\bar{\boldsymbol{\xi}} + \bar{\boldsymbol{\zeta}}\| \leq \epsilon$. Note that $\bar{\boldsymbol{\zeta}}$ is a vector in the normal cone of $\mathcal{Z}$ at $\bar{\mathbf{z}}$ so that $\langle \bar{\boldsymbol{\zeta}}, \bar{\mathbf{z}} - \mathbf{z}\rangle \geq 0$ for any $\mathbf{z} \in \mathcal{Z}$. Hence, we can easily show that

$$\max_{\mathbf{z}\in\mathcal{Z}} \langle \bar{\boldsymbol{\xi}}, \bar{\mathbf{z}} - \mathbf{z}\rangle \leq \max_{\mathbf{z}\in\mathcal{Z}} \langle \bar{\boldsymbol{\xi}} + \bar{\boldsymbol{\zeta}}, \bar{\mathbf{z}} - \mathbf{z}\rangle \leq \|\bar{\boldsymbol{\xi}} + \bar{\boldsymbol{\zeta}}\| D \leq \epsilon D.$$

This means an ϵ-stationary solution is an ϵD-gap solution for the corresponding SVI. However, the reversed direction is not true. We consider the following problem in $\mathbb{R}^2 \times \mathbb{R}$ $\min_{x_1^2+x_2^2\leq r^2} \max_{y\in[-r,r]} x_1$ where the objective function is also viewed as a (constant) function of y. Consider the solution $\bar{\mathbf{z}} = (\bar{x}_1, \bar{x}_2, \bar{y}) = (0, r, 0)$ which is on the boundary of the feasible region and corresponds to $F(\bar{\mathbf{z}}) + 1_{\mathcal{Z}}(\bar{\mathbf{z}}) = \{\boldsymbol{\xi} \in \mathbb{R}^3 | \xi_1 = 1, \xi_2 \geq 0, \xi_3 = 0\}$. As a result, $\text{dist}^2(0, \partial(f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + 1_{\mathcal{Z}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})) = 1$ but, because $F(\bar{\mathbf{z}}) = (1, 0, 0)$, we have $\max_{\mathbf{z}\in\mathcal{Z}} \langle F(\bar{\mathbf{z}}), \bar{\mathbf{z}} - \mathbf{z}\rangle = r$. Then, if r is small, we have $\mathbf{z}$ satisfying (32). This means (34) is stronger than (32).

H.1 Proof of Lemma 20

Proof [Proof of Lemma 20] Since $\bar{\mathbf{w}}$ is the solution to SVI($F^{\gamma}_{\mathbf{w}}, \mathcal{Z}$), there exists $\bar{\boldsymbol{\xi}} \in F(\bar{\mathbf{w}})$ such that

$$\left\langle \bar{\boldsymbol{\xi}} + \frac{1}{\gamma}(\bar{\mathbf{w}} - \mathbf{w}), \mathbf{z} - \bar{\mathbf{w}} \right\rangle \geq 0,$$

for any $\mathbf{z} \in \mathcal{Z}$. The conclusion is proved by reorganizing terms and using the fact that $\|\mathbf{z} - \bar{\mathbf{w}}\| \leq D$. ■

H.2 Proof of Proposition 21

Consider the first statement. Suppose $\mathbf{w} \in \mathcal{Z}$ is a nearly $(\epsilon, c\epsilon)$-gap solution SVI($F, \mathcal{Z}$). By its definition, there exists $\widehat{\mathbf{w}} \in \mathcal{Z}$ such that $\|\mathbf{w} - \widehat{\mathbf{w}}\| \leq c\epsilon$ and $\max_{\mathbf{z}\in\mathcal{Z}}\langle F(\widehat{\mathbf{w}}), \widehat{\mathbf{w}} - \mathbf{z}\rangle \leq \epsilon$. Since $F(\mathbf{z})$ is Lipschitz continuous and $\mathcal{Z}$ is compact, M is finite. Then for any $\mathbf{z} \in \mathcal{Z}$ we have

$$\begin{aligned}
\langle F(\mathbf{w}), \mathbf{w} - \mathbf{z}\rangle &\leq \langle F(\widehat{\mathbf{w}}), \widehat{\mathbf{w}} - \mathbf{z}\rangle + \langle F(\mathbf{w}) - F(\widehat{\mathbf{w}}), \widehat{\mathbf{w}} - \mathbf{z}\rangle + \langle F(\mathbf{w}), \mathbf{w} - \widehat{\mathbf{w}}\rangle \\
&\leq \epsilon + L\|\mathbf{w} - \widehat{\mathbf{w}}\|\|\widehat{\mathbf{w}} - \mathbf{z}\| + M\|\mathbf{w} - \widehat{\mathbf{w}}\| \\
&\leq (1 + LcD + Mc)\,\epsilon
\end{aligned}$$

In light of the above discussion, let us compare the result of applying the extragradient algorithm analyzed in (Dang and Lan, 2015) to the min-max problem (1) when $f(\mathbf{x}, \mathbf{y})$ is smooth both in $\mathbf{x}$ and $\mathbf{y}$ such that the corresponding $F(\mathbf{z})$ is Lipschitz continuous. Their result is that finding a $\bar{\mathbf{z}}$ such that $\max_{\mathbf{z}\in\mathcal{Z}}\langle F(\bar{\mathbf{z}}), \bar{\mathbf{z}} - \mathbf{z}\rangle \leq \epsilon$ requires a complexity of $O(1/\epsilon^2)$. According to Proposition 21, this implies that for finding a nearly ϵ-stationary solution to the min-max problem, the complexity of the extragradient algorithm analyzed in (Dang and Lan, 2015) is $O(1/\epsilon^4)$. In contrast, our complexity in Corollary 23 is $O(\log(1/\epsilon)/\epsilon^2)$. It is worth mentioning that it is unclear whether an improved analysis of the extragradient method as in (Dang and Lan, 2015) can have a better complexity than $O(1/\epsilon^4)$ for finding a nearly ϵ-stationary solution for a min-max saddle-point problem.

References

Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 89–97, 2017.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 699–707, 2016. URL `http://jmlr.org/proceedings/papers/v48/allen-zhua16.html`.

Elisabetta Allevi, Adriana Gnudi, and Igor V Konnov. The proximal point method for nonmonotone variational inequalities. *Mathematical Methods of Operations Research*, 63 (3):553–565, 2006.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 2011.

T. Q. Bao and P. Q. Khanh. A projection-type algorithm for pseudomonotone nonlipschitzian multivalued variational inequalities. *Nonconvex Optim. Appl.*, 77:113?129, 2005.

Regina S. Burachik and R. Diaz Millan. A projection algorithm for non-monotone variational inequalities. *CoRR*, abs/1609.09569, 2016.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, May 2011. ISSN 0924-9907. doi: 10.1007/s10851-010-0251-1. URL `http://dx.doi.org/10.1007/s10851-010-0251-1`.

Robert S. Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 4705–4714. 2017.

Zaiyi Chen and Tianbao Yang. A variance reduction method for non-convex optimization with improved convergence under large condition number. *CoRR*, abs/1809.06754, 2018.

Zaiyi Chen, Tianbao Yang, Jinfeng Yi, Bowen Zhou, and Enhong Chen. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *CoRR*, /abs/1808.06296, 2018.

A. Cherukuri, B. Gharesifard, and J. CortÃ©s. Saddle-point dynamics: Conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1): 486–511, 2017. doi: 10.1137/15M1026924.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1133–1142, 2018.

Cong D. Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Comp. Opt. and Appl.*, 60(2):277–310, 2015.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in neural information processing systems*, pages –, 2018.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *CoRR*, abs/1711.00141, 2017.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *arXiv preprint arXiv:1803.06523*, 2018a.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *CoRR*, /abs/1802.02988, 2018b.

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *arXiv preprint arXiv:1707.03505*, 2017.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *CoRR*, 2017.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

Gauthier Gidel, Hugo Berard, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *CoRR*, abs/1802.10551, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Paulina Grnarova, Kfir Y. Levy, Aurélien Lucchi, Thomas Hofmann, and Andreas Krause. An online learning approach to generative adversarial networks. *CoRR*, abs/1706.03269, 2017.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 5769–5779, 2017.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3):161–220, 1990.

Philip Hartman and Guido Stampacchia. On some non-linear elliptic differential-functional equations. *Acta mathematica*, 115(1):271–310, 1966.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 6629–6640, 2017.

Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of ACM*, 60(6):45:1–45:39, November 2013. ISSN 0004-5411.

A. N. Iusem, A. Jofré, Roberto Imbuzeiro Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2698–2707, 2018.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Guanghui Lan and Yu Yang. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *CoRR*, abs/1805.05411, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems 30*, pages 597–607. 2017.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *CoRR*, abs/1807.02629, 2018.

George J Minty et al. Monotone (nonlinear) operators in hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 5591–5600, 2017.

A. Nemirovski and D. Yudin. On cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathematics Doklady*, 19: 341–362, 1978.

Arkadi Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. on Optimization*, 15(1):229–251, January 2005.

Arkadi Nemirovski. *Information-based Complexity of Convex Programming*. 2015.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete and Continuous Dynamical Systems - A*, 31:1383, 2011.

Balamurugan Palaniappan and Francis R. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1408–1416, 2016.

Qi Qian, Shenghuo Zhu, Jiasheng Tang, Rong Jin, Baigui Sun, and Hao Li. Robust optimization over multiple domains. *CoRR*, abs/1805.07588, 2018.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *CoRR*, abs/1810.02060, 2018.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczós, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pages 314–323. JMLR.org, 2016a.

Sashank J. Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Fast incremental method for smooth nonconvex optimization. In *55th IEEE Conference on Decision and Control (CDC)*, pages 1971–1977, 2016b.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. on Control and Optimization*, 14(5), 1976.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Ohad Shamir. Can we find near-approximately-stationary points of nonsmooth nonconvex functions? *arXiv preprint arXiv:2002.11962*, 2020.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifiable distributional robustness with principled adversarial training. *CoRR*, abs/1710.10571, 2017.

Mikhail V. Solodov and Benar Fux Svaiter. A new projection method for variational inequality problems. *SIAM J. CONTROL OPTIM*, 1999.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pages 12659–12670, 2019.

Y. J. Wang, N. H. Xiu, and C. Y. Wang. Unified framework of extragradient-type methods for pseudomonotone variational inequalities. *Journal of Optimization Theory and Applications*, 111(3):641–656, Dec 2001.

Yi Xu, Shenghuo Zhu, Sen Yang, Chi Zhang, Rong Jin, and Tianbao Yang. Learning with non-convex truncated losses by SGD. *CoRR*, abs/1805.07880, 2018. URL `http://arxiv.org/abs/1805.07880`.

Tianbao Yang, Mehrdad Mahdavi, Rong Jin, and Shenghuo Zhu. An efficient primal dual prox method for non-smooth optimization. *Machine Learning*, 98(3):369–406, 2015.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. volume abs/1604.03257, 2016.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Ali Jadbabaie, and Suvrit Sra. Complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020.

Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.