On Gradient Descent Ascent for Nonconvex-Concave Minimax Problems

Tianyi Lin [1] Chi Jin [2] Michael. I. Jordan [3]

Abstract

We consider nonconvex-concave minimax problems, $\min_{\mathbf{x}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})$ where f is nonconvex in $\mathbf{x}$ but concave in $\mathbf{y}$ and $\mathcal{Y}$ is a convex and bounded set. One of the most popular algorithms for solving this problem is the celebrated gradient descent ascent (GDA) algorithm, which has been widely used in machine learning, control theory and economics. Despite the extensive convergence results for the convex-concave setting, GDA with equal stepsize can converge to limit cycles or even diverge in a general setting. In this paper, we present the complexity results on two-time-scale GDA for solving nonconvex-concave minimax problems, showing that the algorithm can find a stationary point of the function $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot,\mathbf{y})$ efficiently. To the best our knowledge, this is the first nonasymptotic analysis for two-time-scale GDA in this setting, shedding light on its superior practical performance in training generative adversarial networks (GANs) and other real applications.

1. Introduction

We consider the following smooth minimax optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}), \tag{1}$$

where $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is nonconvex in $\mathbf{x}$ but concave in $\mathbf{y}$ and where $\mathcal{Y}$ is a convex set. Since von Neumann's seminal work (Neumann, 1928), the problem of finding the solution to problem (1) has been a major focus of research in mathematics, economics and computer science (Basar & Olsder, 1999; Nisan et al., 2007; Von Neumann & Morgenstern, 2007). In recent years, minimax optimization theory has begun to see applications in machine learning, with examples including generative adversarial networks (GANs) (Goodfellow et al., 2014), statistics (Xu et al., 2009; Abadeh et al., 2015), online learning (Cesa-Bianchi & Lugosi, 2006), deep learning (Sinha et al., 2018) and distributed computing (Shamma, 2008; Mateos et al., 2010). Moreover, there is increasing awareness that machine-learning systems are embedded in real-world settings involving scarcity or competition that impose game-theoretic constraints (Jordan, 2018).

One of the simplest candidates for solving problem (1) is the natural generalization of gradient descent (GD) known as *gradient descent ascent* (GDA). At each iteration, this algorithm performs gradient descent over the variable $\mathbf{x}$ with the stepsize $\eta_{\mathbf{x}}$ and gradient ascent over the variable $\mathbf{y}$ with the stepsize $\eta_{\mathbf{y}}$. On the positive side, when the objective function f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, there is a vast literature establishing asymptotic and nonasymptotic convergence for the average iterates generated by GDA with the equal stepsizes ($\eta_{\mathbf{x}} = \eta_{\mathbf{y}}$); (see, e.g., Korpelevich, 1976; Chen & Rockafellar, 1997; Nedić & Ozdaglar, 2009; Nemirovski, 2004; Du & Hu, 2018). Local linear convergence can also be shown under the additional assumption that f is locally strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$ (Cherukuri et al., 2017; Adolphs et al., 2018; Liang & Stokes, 2018). However, there has been no shortage of research highlighting the fact that in a general setting GDA with equal stepsizes can converge to limit cycles or even diverge (Benaım & Hirsch, 1999; Hommes & Ochea, 2012; Mertikopoulos et al., 2018).

Recent research has focused on alternative gradient-based algorithms that have guarantees beyond the convex-concave setting (Daskalakis et al., 2017; Heusel et al., 2017; Mertikopoulos et al., 2019; Mazumdar et al., 2019). Two-time-scale GDA (Heusel et al., 2017) has been particularly popular. This algorithm, which involves unequal stepsizes ($\eta_{\mathbf{x}} \neq \eta_{\mathbf{y}}$), has been shown to empirically to alleviate the issues of limit circles and it has theoretical support in terms of local asymptotic convergence to Nash equilibria; (Heusel et al., 2017, Theorem 2).

This asymptotic result stops short of providing an understanding of algorithmic efficiency, and it would be desirable to provide a stronger, nonasymptotic, theoretical convergence rate for two-time-scale GDA in a general setting. In particular, the following general structure arises in many applications: $f(\mathbf{x},\cdot)$ is concave for any $\mathbf{x}$ and $\mathcal{Y}$ is a bounded

[*] Equal contribution [1] Department of Industrial Engineering and Operations Research, UC Berkeley [2] Department of Electrical Engineering, Princeton University [3] Department of Statistics and Electrical Engineering and Computer Science, UC Berkeley. Correspondence to: Tianyi Lin <darren_lin@berkeley.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Table 1. The gradient complexity of all algorithms for nonconvex-(strongly)-concave minimax problems. ϵ is a tolerance and $\kappa > 0$ is a condition number. The result denoted by * refers to the complexity bound after translating from ϵ-stationary point of f to our optimality measure; see Propositions 4.11 and 4.12. The result denoted by $^\circ$ is not presented explicitly but easily derived by standard arguments.

	Nonconvex-Strongly-Concave		Nonconvex-Concave		Simplicity
	Deterministic	Stochastic	Deterministic	Stochastic	
Jin et al. (2019)	$\tilde{O}\left(\kappa^2\epsilon^{-2}\right)^{\circ}$	$\tilde{O}\left(\kappa^3\epsilon^{-4}\right)$	$O(\epsilon^{-6})$	$O(\epsilon^{-8})^{\circ}$	Double-loop
Rafique et al. (2018)	$\tilde{O}(\kappa^2\epsilon^{-2})$	$\tilde{O}(\kappa^3\epsilon^{-4})$	$\tilde{O}(\epsilon^{-6})$	$\tilde{O}(\epsilon^{-6})$	Double-loop
Nouiehed et al. (2019)	$\tilde{O}(\kappa^4\epsilon^{-2})^{\star,\circ}$	–	$O(\epsilon^{-7})^{\star}$	–	Double-loop
Thekumparampil et al. (2019)	–	–	$\tilde{O}(\epsilon^{-3})$	–	Triple-loop
Kong & Monteiro (2019)	–	–	$\tilde{O}(\epsilon^{-3})$	–	Triple-loop
Lu et al. (2019)	$O(\kappa^4\epsilon^{-2})^{\star}$	–	$O(\epsilon^{-8})^{\star}$	–	Single-loop
This paper	$O(\kappa^2\epsilon^{-2})$	$O(\kappa^3\epsilon^{-4})$	$O(\epsilon^{-6})$	$O(\epsilon^{-8})$	Single-loop

set. Two typical examples are the training of a neural network which is robust to adversarial examples (Madry et al., 2017) and the learning of a robust classifier from multiple distributions (Sinha et al., 2018). Both of these schemes can be posed as nonconvex-concave minimax problems. Based on this observation, it is natural to ask the question: Are two-time-scale GDA and stochastic GDA (SGDA) provably efficient for nonconvex-concave minimax problems?

Our results: This paper presents an affirmative answer to this question, providing nonasymptotic complexity results for two-time scale GDA and SGDA in two settings. In the nonconvex-strongly-concave setting, two-time scale GDA and SGDA require $O(\kappa^2\epsilon^{-2})$ gradient evaluations and $O(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations, respectively, to return an ϵ-stationary point of the function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ where $\kappa > 0$ is a condition number. In the nonconvex-concave setting, two-time scale GDA and SGDA require $O(\epsilon^{-6})$ gradient evaluations and $O(\epsilon^{-8})$ stochastic gradient evaluations.

Main techniques: To motivate the proof ideas for analyzing two-time scale GDA and SGDA, it is useful to contrast our work with some of the strongest existing convergence analyses for nonconvex-concave problems. In particular, Jin et al. (2019) and Nouiehed et al. (2019) have provided complexity results for algorithms that have a nested-loop structure. Specifically, GDmax and multistep GDA are algorithms in which the outer loop can be interpreted as an inexact gradient descent on a nonconvex function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ while the inner loop provides an approximate solution to the maximization problem $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ for a given $\mathbf{x} \in \mathbb{R}^m$. Strong convergence results are obtained when accelerated gradient ascent is used in the maximization problem.

Compared to GDmax and multistep GDA, two-time scale GDA and SGDA are harder to analyze. Indeed, $\mathbf{y}_t$ is not necessarily guaranteed to be close to $\mathbf{y}^\star(\mathbf{x}_t)$ at each iteration and thus it is unclear that $\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$ might a reasonable descent direction. To overcome this difficulty, we develop a new technique which analyzes the concave optimization with a slowly changing objective function. This is the main technical contribution of this paper.

Notation. We use bold lower-case letters to denote vectors and caligraphic upper-case letter to denote sets. We use $\|\cdot\|$ to denote the ℓ_2-norm of vectors and spectral norm of matrices. For a function $f : \mathbb{R}^n \to \mathbb{R}$, $\partial f(\mathbf{z})$ denotes the subdifferential of f at $\mathbf{z}$. If f is differentiable, $\partial f(\mathbf{z}) = \{\nabla f(\mathbf{z})\}$ where $\nabla f(\mathbf{z})$ denotes the gradient of f at $\mathbf{z}$ and $\nabla_{\mathbf{x}} f(\mathbf{z})$ denotes the partial gradient of f with respect to $\mathbf{x}$ at $\mathbf{z}$. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the largest and smallest eigenvalue of A denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$.

2. Related Work

Convex-concave setting. Historically, an early concrete instantiation of problem (1) involved computing a pair of probability vectors $(\mathbf{x}, \mathbf{y})$, or equivalently solving $\min_{\mathbf{x}\in\Delta^m} \max_{\mathbf{y}\in\Delta^n} \mathbf{x}^\top A\mathbf{y}$ for a matrix $A \in \mathbb{R}^{m\times n}$ and probability simplices Δ^m and Δ^n. This bilinear minimax problem together with von Neumann's minimax theorem (Neumann, 1928) was a cornerstone in the development of game theory. A general algorithm scheme was developed for solving this problem in which the min and max players each run a simple learning procedure in tandem (Robinson, 1951). Sion (1958) generalized von Neumann's result from bilinear games to general convex-concave games, $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}} \min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$, and triggered a line of algorithmic research on convex-concave minimax optimization in both continuous time (Kose, 1956; Cherukuri et al., 2017) and discrete time (Uzawa, 1958; Golshtein, 1974; Korpelevich, 1976; Nemirovski, 2004; Nedić & Ozdaglar, 2009; Mokhtari et al., 2019b;a; Azizian et al., 2019). It is well known that GDA finds an ϵ-approximate stationary point within $O(\kappa^2 \log(1/\epsilon))$ iterations for strongly-convex-strongly-concave problems, and $O(\epsilon^{-2})$ iterations with decaying stepsize for convex-concave games (Nedić & Ozdaglar, 2009; Nemirovski, 2004).

Nonconvex-concave setting. Nonconvex-concave minimax problems appear to be a class of tractable problems in the form of problem (1) and have emerged as a focus in optimization and machine learning (Namkoong & Duchi, 2016; Sinha et al., 2018; Rafique et al., 2018; Sanjabi et al., 2018; Grnarova et al., 2018; Lu et al., 2019; Nouiehed et al., 2019; Thekumparampil et al., 2019; Kong & Monteiro, 2019); see Table 1 for a comprehensive overview. We also wish to highlight the work of Grnarova et al. (2018), who proposed a variant of GDA for nonconvex-concave problem and the work of Sinha et al. (2018); Sanjabi et al. (2018), who studied a class of inexact nonconvex SGD algorithms that can be categorized as variants of SGDmax for nonconvex-strongly-concave problem. Jin et al. (2019) analyzed the GDmax algorithm for nonconvex-concave problem and provided nonasymptotic convergence results.

Rafique et al. (2018) proposed “proximally guided stochastic mirror descent” and “variance reduced gradient” algorithms (PGSMD/PGSVRG) and proved that these algorithms find an approximate stationary point of $\Phi(\cdot) := \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$. However, PGSMD/PGSVRG are nested-loop algorithms and convergence results were established only in the special case where $f(\mathbf{x}, \cdot)$ is a linear function (Rafique et al., 2018, Assumption 2 D.2). Nouiehed et al. (2019) developed a multistep GDA (MGDA) algorithm by incorporating accelerated gradient ascent as the subroutine at each iteration. This algorithm provably finds an approximate stationary point of $f(\cdot, \cdot)$ for nonconvex-concave problems with the fast rate of $O(\epsilon^{-3.5})$. Very recently, Thekumparampil et al. (2019) proposed a proximal dual implicit accelerated gradient (ProxDIAG) algorithm for nonconvex-concave problems and proved that the algorithm find an approximate stationary point of $\Phi(\cdot)$ with the rate of $O(\epsilon^{-3})$. This complexity result is also achieved by an inexact proximal point algorithm (Kong & Monteiro, 2019). All of these algorithms are, however, nested-loop algorithms and thus relatively complicated to implement. One would like to know whether the nested-loop structure is necessary or whether GDA, a single-loop algorithm, can be guaranteed to converge in the nonconvex-(strongly)-concave setting.

The most closest work to ours is Lu et al. (2019), where a single-loop HiBSA algorithm for nonconvex-(strongly)-concave problems is proposed with theoretical guarantees under a different notion of optimality. However, their analysis requires some restrictive assumptions; e.g., that $f(\cdot, \cdot)$ is lower bounded. We only require that $\max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ is lower bounded. An example which meets our conditions and not those of Lu et al. (2019) is $\min_{\mathbf{x}\in\mathbb{R}} \max_{\mathbf{y}\in[-1,1]} \mathbf{x}^\top \mathbf{y}$. Our less-restrictive assumptions make the problem more challenging and our technique is accordingly fundamentally difference from theirs.

Nonconvex-nonconcave setting. During the past decade, the study of nonconvex-nonconcave minimax problems has become a central topic in machine learning, inspired in part by the advent of generative adversarial networks (Goodfellow et al., 2014) and adversarial learning (Madry et al., 2017; Namkoong & Duchi, 2016; Sinha et al., 2018). Most recent work aims at defining a notion of goodness or the development of new procedures for reducing oscillations (Daskalakis & Panageas, 2018b; Adolphs et al., 2018; Mazumdar et al., 2019) and speeding up the convergence of gradient dynamics (Heusel et al., 2017; Balduzzi et al., 2018; Mertikopoulos et al., 2019; Lin et al., 2018). Daskalakis & Panageas (2018b) study minimax optimization (or zero-sum games) and show that the stable limit points of GDA are not necessarily Nash equilibria. Adolphs et al. (2018) and Mazumdar et al. (2019) propose Hessian-based algorithms whose stable fixed points are exactly Nash equilibria. On the other hand, Balduzzi et al. (2018) develop a new symplectic gradient adjustment (SGA) algorithm for finding stable fixed points in potential games and Hamiltonian games. Heusel et al. (2017) propose two-time-scale GDA and show that Nash equilibria are stable fixed points of the continuous limit of two-time-scale GDA under certain strong conditions. All of these convergence results are either local or asymptotic and not extend to cover our results in a nonconvex-concave setting. Very recently, Mertikopoulos et al. (2019); Lin et al. (2018) provide nonasymptotic guarantees for a special class of nonconvex-nonconcave minimax problems under variational stability and the Minty condition. However, while both of these two conditions must hold in convex-concave setting, they do not necessarily hold in nonconvex-(strongly)-concave problem.

Online learning setting. From the online learning perspective, there is difference in no-regret property of different algorithms. For example, the extragradient algorithm (Mertikopoulos et al., 2019) is not no-regret, while the optimistic algorithm (Daskalakis & Panageas, 2018a) is a no-regret algorithm. In comparing limit behavior of zero-sum game dynamics, Bailey & Piliouras (2018) showed that the multiplicative weights update has similar property as GDA and specified the necessity of introducing the optimistic algorithms to study the last-iterate convergence.

3. Preliminaries

We recall basic definitions for smooth functions.

Definition 3.1 *A function f is **L-Lipschitz** if for $\forall \mathbf{x}, \mathbf{x}'$, we have $\|f(\mathbf{x}) - f(\mathbf{x}')\| \leq L \|\mathbf{x} - \mathbf{x}'\|$.*

Definition 3.2 *A function f is **ℓ-smooth** if for $\forall \mathbf{x}, \mathbf{x}'$, we have $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq \ell \|\mathbf{x} - \mathbf{x}'\|$.*

Recall that the minimax problem (1) is equivalent to minimizing a function $\Phi(\cdot) = \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$. For nonconvex-concave minimax problems in which $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$, the maximization problem $\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ can be solved efficiently and provides useful information about Φ. However, it is still NP hard to find the global minimum of Φ in general since Φ is nonconvex.

Objectives in this paper. We start by defining local surrogate for the global minimum of Φ. A common surrogate in nonconvex optimization is the notion of stationarity, which is appropriate if Φ is differentiable.

Definition 3.3 *A point $\mathbf{x}$ is an* ϵ-stationary point *($\epsilon \geq 0$) of a differentiable function Φ if $\|\nabla \Phi(\mathbf{x})\| \leq \epsilon$. If $\epsilon = 0$, then $\mathbf{x}$ is a stationary point.*

Definition 3.3 is sufficient for nonconvex-strongly-concave minimax problem since $\Phi(\cdot) = \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ is differentiable in that setting. In contrast, a function Φ is not necessarily differentiable for general nonconvex-concave minimax problem even if f is Lipschitz and smooth. A weaker condition that we make use of is the following.

Definition 3.4 *A function Φ is* ℓ-weakly convex *if a function $\Phi(\cdot) + (\ell/2)\|\cdot\|^2$ is convex.*

For a ℓ-weakly convex function Φ, the subdifferential $\partial \Phi$ is uniquely determined by the subdifferential of $\Phi + (\ell/2)\|\cdot\|^2$. Thus, a naive measure of approximate stationarity can be defined as a point $\mathbf{x} \in \mathbb{R}^m$ such that at least one subgradient is small: $\min_{\xi \in \partial \Phi(\mathbf{x})} \|\xi\| \leq \epsilon$. However, this notion of stationarity can be very restrictive when optimizing nonsmooth functions. For example, when $\Phi(\cdot) = |\cdot|$ is a one-dimensional function, an ϵ-stationary point is zero for all $\epsilon \in [0, 1)$. This means that finding an approximate stationary point under this notion is as difficult as solving the problem exactly. To alleviate this issue, Davis & Drusvyatskiy (2019) propose an alternative notion of stationarity based on the Moreau envelope. This has become recognized as standard for optimizing a weakly convex function.

Definition 3.5 *A function $\Phi_\lambda : \mathbb{R}^m \to \mathbb{R}$ is the Moreau envelope of Φ with a positive parameter $\lambda > 0$ if $\Phi_\lambda(\mathbf{x}) = \min_{\mathbf{w}} \Phi(\mathbf{w}) + (1/2\lambda)\|\mathbf{w} - \mathbf{x}\|^2$ for each $\mathbf{x} \in \mathbb{R}^m$.*

Algorithm 1 Two-Time-Scale GDA

```
Input: $(\mathbf{x}_0, \mathbf{y}_0)$, stepsizes $(\eta_\mathbf{x}, \eta_\mathbf{y})$.
for $t = 1, 2, \ldots, T$ do
  $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_\mathbf{x} \nabla_\mathbf{x} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$,
  $\mathbf{y}_t \leftarrow \mathcal{P}_\mathcal{Y} \left(\mathbf{y}_{t-1} + \eta_\mathbf{y} \nabla_\mathbf{y} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right)$.
Randomly draw $\hat{\mathbf{x}}$ from $\{\mathbf{x}_t\}_{t=1}^T$ at uniform.
Return: $\hat{\mathbf{x}}$.
```

Lemma 3.6 *If a function f is ℓ-smooth and $\mathcal{Y}$ is bounded, the Moreau envelope $\Phi_{1/2\ell}$ of $\Phi(\cdot) = \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ is differentiable, ℓ-smooth and ℓ-strongly convex.*

Thus, an alternative measure of approximate stationarity of a function $\Phi(\cdot) = \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ can be defined as a point $\mathbf{x} \in \mathbb{R}^m$ such that the norm of the gradient of Moreau envelope is small: $\|\nabla \Phi_{1/2\ell}\| \leq \epsilon$. More generally, we have

Definition 3.7 *A point $\mathbf{x}$ is an ϵ-stationary point ($\epsilon \geq 0$) of a ℓ-weakly convex function Φ if $\|\nabla \Phi_{1/2\ell}(\mathbf{x})\| \leq \epsilon$. If $\epsilon = 0$, then $\mathbf{x}$ is a stationary point.*

Although Definition 3.7 uses the language of Moreau envelopes, it also connects to the function Φ as follows.

Lemma 3.8 *If $\mathbf{x}$ is an ϵ-stationary point of a ℓ-weakly convex function Φ (Definition 3.7), there exists $\hat{\mathbf{x}} \in \mathbb{R}^m$ such that $\min_{\xi \in \partial \Phi(\hat{\mathbf{x}})} \|\xi\| \leq \epsilon$ and $\|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon/2\ell$.*

Lemma 3.8 shows that an ϵ-stationary point defined by Definition 3.7 can be interpreted as the relaxation or surrogate for $\min_{\xi \in \partial \Phi(\mathbf{x})} \|\xi\| \leq \epsilon$. In particular, if a point $\mathbf{x}$ is an ϵ-stationary point of an ℓ-weakly convex function Φ, then $\mathbf{x}$ is close to a point $\hat{\mathbf{x}}$ which has at least one small subgradient.

Remark 3.9 *We remark that our notion of stationarity is natural in real scenarios. Indeed, many applications arising from adversarial learning can be formulated as the minimax problem (1), and, in this setting, $\mathbf{x}$ is the classifier while $\mathbf{y}$ is the adversarial noise for the data. Practitioners are often interested in finding a robust classifier $\mathbf{x}$ instead of recovering the adversarial noise $\mathbf{y}$. Any stationary point of the function $\Phi(\cdot) = \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y})$ corresponds precisely to a robust classifier that achieves better classification error.*

Remark 3.10 *There are also other notions of stationarity based on ∇f are proposed for nonconvex-concave minimax problems in the literature (Lu et al., 2019; Nouiehed et al., 2019). However, as pointed by Thekumparampil et al. (2019), these notions are weaker than that defined in Definition 3.3 and 3.7. For the sake of completeness, we specify the relationship between our notion of stationarity and other notions in Proposition 4.11 and 4.12.*

Algorithm 2 Two-Time-Scale SGDA

```
Input: $(\mathbf{x}_0, \mathbf{y}_0)$, stepsizes $(\eta_\mathbf{x}, \eta_\mathbf{y})$, batch size $M$.
for $t = 1, 2, \ldots, T$ do
    Draw a collection of i.i.d. data samples $\{\xi_i\}_{i=1}^M$.
    $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_\mathbf{x} \left(\frac{1}{M}\sum_{i=1}^M G_\mathbf{x}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right)$.
    $\mathbf{y}_t \leftarrow \mathcal{P}_\mathcal{Y}\left(\mathbf{y}_{t-1} + \eta_\mathbf{y}(\frac{1}{M}\sum_{i=1}^M G_\mathbf{y}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i))\right)$.
Randomly draw $\hat{\mathbf{x}}$ from $\{\mathbf{x}_t\}_{t=1}^T$ at uniform.
Return: $\hat{\mathbf{x}}$.
```

4. Main Results

In this section, we present complexity results for two-time-scale GDA and SGDA in the setting of nonconvex-strongly-concave and nonconvex-concave minimax problems.

The algorithmic schemes that we study are extremely simple and are presented in Algorithm 1 and 2. In particular, each iteration comprises one (stochastic) gradient descent step over $\mathbf{x}$ with the stepsize $\eta_\mathbf{x} > 0$ and one (stochastic) gradient ascent step over $\mathbf{y}$ with the stepsize $\eta_\mathbf{y} > 0$. The choice of stepsizes $\eta_\mathbf{x}$ and $\eta_\mathbf{y}$ is crucial for the algorithms in both theoretical and practical senses. In particular, classical GDA and SGDA assume that $\eta_\mathbf{x} = \eta_\mathbf{y}$, and the last iterate is only known convergent in strongly convex-concave problems (Liang & Stokes, 2018). Even in convex-concave settings (or bilinear settings as special cases), GDA requires the assistance of averaging or other strategy (Daskalakis & Panageas, 2018a) to converge, otherwise, with fixed stepsize, the last iterate will always diverge and hit the constraint boundary eventually (Daskalakis et al., 2017; Mertikopoulos et al., 2018; Daskalakis & Panageas, 2018a). In contrast, two-time-scale GDA and SGDA ($\eta_\mathbf{x} \neq \eta_\mathbf{y}$) were shown to be locally convergent and practical in training GANs (Heusel et al., 2017).

One possible reason for this phenomenon is that the choice of $\eta_\mathbf{x} \neq \eta_\mathbf{y}$ reflects the nonsymmetric nature of nonconvex-(strongly)-concave problems. For sequential problems such as robust learning, where the natural order of min-max is important (i.e., min-max is not equal to max-min), practitioners often prefer faster convergence for the inner max problem. Therefore, it is reasonable for us to choose $\eta_\mathbf{x} \ll \eta_\mathbf{y}$ rather than $\eta_\mathbf{x} = \eta_\mathbf{y}$.

Finally, we make the standard assumption that the oracle $G = (G_\mathbf{x}, G_\mathbf{y})$ is unbiased and has bounded variance.

Assumption 4.1 *The stochastic oracle G satisfies*

$$\begin{aligned}\mathbb{E}[G(\mathbf{x}, \mathbf{y}, \xi) - \nabla f(\mathbf{x}, \mathbf{y}] &= 0, \\ \mathbb{E}[\|G(\mathbf{x}, \mathbf{y}, \xi) - \nabla f(\mathbf{x}, \mathbf{y})\|^2] &\leq \sigma^2.\end{aligned}$$

4.1. Nonconvex-strongly-concave minimax problems

In this subsection, we present the complexity results for two-time-scale GDA and SGDA in the setting of nonconvex-strongly-concave minimax problems. The following assumption is made throughout this subsection.

Assumption 4.2 *The objective function and constraint set ($f : \mathbb{R}^{m+n} \rightarrow \mathbb{R}$, $\mathcal{Y} \subseteq \mathbb{R}^n$) satisfy*

1. *f is ℓ-smooth and $f(\mathbf{x}, \cdot)$ is μ-strongly concave.*
2. *$\mathcal{Y}$ is a convex and bounded set with a diameter $D \geq 0$.*

Let $\kappa = \ell/\mu$ denote the condition number and define

$$\Phi(\cdot) \;=\; \max_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y}), \quad \mathbf{y}^\star(\cdot) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} f(\cdot, \mathbf{y}).$$

We present a technical lemma on the structure of the function Φ in the nonconvex-strongly-concave setting.

Lemma 4.3 *Under Assumption 4.2, $\Phi(\cdot)$ is $(\ell + \kappa\ell)$-smooth with $\nabla\Phi(\cdot) = \nabla_\mathbf{x} f(\cdot, \mathbf{y}^\star(\cdot))$. Also, $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz.*

Since Φ is differentiable, the notion of stationarity in Definition 3.3 is our target given only access to the (stochastic) gradient of f. Denote $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_\mathbf{x} \Phi(\mathbf{x})$, we proceed to provide theoretical guarantees for two-time-scale GDA and SGDA algorithms.

Theorem 4.4 (GDA) *Under Assumption 4.2 and letting the stepsizes be chosen as $\eta_\mathbf{x} = \Theta(1/\kappa^2\ell)$ and $\eta_\mathbf{y} = \Theta(1/\ell)$, the iteration complexity (also the gradient complexity) of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right).$$

Theorem 4.5 (SGDA) *Under Assumption 4.1 and 4.2 and letting the stepsizes $\eta_\mathbf{x}, \eta_\mathbf{y}$ be chosen as the same in Theorem 4.4 with the batch size $M = \Theta(\max\{1, \kappa\sigma^2\epsilon^{-2}\})$, the iteration complexity of Algorithm 2 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right),$$

which gives the total stochastic gradient complexity:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2} \max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

We make several remarks.

First, two-time-scale GDA and SGDA are guaranteed to find an ϵ-stationary point of $\Phi(\cdot)$ within $O(\kappa^2\epsilon^{-2})$ gradient evaluations and $O(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations, respectively. The ratio of stepsizes $\eta_{\mathbf{y}}/\eta_{\mathbf{x}}$ is required to be $\Theta(\kappa^2)$ due to the nonsymmetric nature of our problem (min-max is not equal to max-min). The quantity $O(\kappa^2)$ reflects an efficiency trade-off in the algorithm.

Furthermore, both of the algorithms are only guaranteed to visit an ϵ-stationary point within a certain number of iterations and return $\hat{\mathbf{x}}$ which is drawn from $\{\mathbf{x}_t\}_{t=1}^T$ at uniform. This does not mean that the last iterate $\mathbf{x}_T$ is the ϵ-stationary point. Such a scheme and convergence result are standard in nonconvex optimization for GD or SGD to find stationary points. In practice, one usually returns the iterate when the learning curve stops changing significantly.

Finally, the minibatch size $M = \Theta(\epsilon^{-2})$ is necessary for the convergence property of two-time-scale SGDA. Even though our proof technique can be extended to the purely stochastic setting ($M = 1$), the complexity result becomes worse, i.e., $O(\kappa^3\epsilon^{-5})$. It remains open whether this gap can be closed or not and we leave it as future work.

4.2. Nonconvex-concave minimax problems

In this subsection, we present the complexity results for two-time-scale GDA and SGDA in the nonconvex-concave minimax setting. The following assumption is made throughout this subsection.

Assumption 4.6 *The objective function and constraint set, $(f : \mathbb{R}^{m+n} \rightarrow \mathbb{R},\ \mathcal{Y} \subset \mathbb{R}^n)$ satisfy*

1. *f is ℓ-smooth and $f(\cdot, \mathbf{y})$ is L-Lipschitz for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathbb{R}^m$.*

2. *$\mathcal{Y}$ is a convex and bounded set with a diameter $D \geq 0$.*

Since $f(\mathbf{x}, \cdot)$ is merely concave for each $\mathbf{x} \in \mathbb{R}^m$, the function $\Phi(\cdot) = \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ is possibly not differentiable. Fortunately, the following structural lemma shows that Φ is ℓ-weakly convex and L-Lipschitz.

Lemma 4.7 *Under Assumption 4.6, $\Phi(\cdot)$ is ℓ-weakly convex and L-Lipschitz with $\nabla_{\mathbf{x}} f(\cdot, \mathbf{y}^\star(\cdot)) \in \partial\Phi(\cdot)$ where $\mathbf{y}^\star(\cdot) \in \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$.*

Since Φ is ℓ-weakly convex, the notion of stationarity in Definition 3.7 is our target given only access to the (stochastic) gradient of f. Denote $\widehat{\Delta}_\Phi = \Phi_{1/2\ell}(\mathbf{x}_0) - \min_{\mathbf{x}} \Phi_{1/2\ell}(\mathbf{x})$ and $\widehat{\Delta}_0 = \Phi(\mathbf{x}_0) - f(\mathbf{x}_0, \mathbf{y}_0)$, we present complexity results for two-time-scale GDA and SGDA algorithms.

Theorem 4.8 (GDA) *Under Assumption 4.6 and letting the step sizes be chosen as $\eta_{\mathbf{x}} = \Theta(\epsilon^4/(\ell^3L^2D^2))$ and $\eta_{\mathbf{y}} = \Theta(1/\ell)$, the iteration complexity (also the gradient complexity) of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\ell^3L^2D^2\widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3D^2\widehat{\Delta}_0}{\epsilon^4}\right).$$

Theorem 4.9 (SGDA) *Under Assumption 4.1 and 4.6 and letting the step sizes be chosen as $\eta_{\mathbf{x}} = \Theta(\epsilon^4/(\ell^3D^2(L^2 + \sigma^2)))$ and $\eta_{\mathbf{y}} = \Theta(\epsilon^2/\ell\sigma^2)$ with the batchsize $M = 1$, the iteration complexity (also the stochastic gradient complexity) of Algorithm 2 to return an ϵ-stationary point is bounded by*

$$\mathcal{O}\left(\left(\frac{\ell^3\left(L^2+\sigma^2\right)D^2\widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3D^2\Delta_0}{\epsilon^4}\right)\max\left\{1, \frac{\sigma^2}{\epsilon^2}\right\}\right).$$

We make several additional remarks. First, two-time-scale GDA and SGDA are guaranteed to find an ϵ-stationary point in terms of Moreau envelopes within $O(\epsilon^{-6})$ gradient evaluations and $O(\epsilon^{-8})$ stochastic gradient evaluations, respectively. The ratio of stepsizes $\eta_{\mathbf{y}}/\eta_{\mathbf{x}}$ is required to be $\Theta(1/\epsilon^4)$ and this quantity reflects an efficiency trade-off in the algorithm. Furthermore, similar arguments as in Section 4.1 hold for the output of the algorithms here. Finally, the minibatch size $M = 1$ is allowed in Theorem 4.9, which is different from the result in Theorem 4.5.

4.3. Relationship between the stationarity notions

We provide additional technical results on the relationship between our notions of stationarity and other notions based on ∇f in the literature (Lu et al., 2019; Nouiehed et al., 2019). In particular, we show that two notions can be translated in both directions with extra computational cost.

Definition 4.10 *A pair of points $(\mathbf{x}, \mathbf{y})$ is an ϵ-stationary point ($\epsilon \geq 0$) of a differentiable function Φ if*

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\| &\leq \epsilon, \\ \|\mathcal{P}_{\mathcal{Y}}(\mathbf{y} + (1/\ell)\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})) - \mathbf{y}\| &\leq \epsilon/\ell.\end{aligned}$$

We present our results in the following two propositions.

Proposition 4.11 *Under Assumption 4.2, if a point $\hat{\mathbf{x}}$ is an ϵ-stationary point in terms of Definition 3.3, an $O(\epsilon)$-stationary point $(\mathbf{x}', \mathbf{y}')$ in terms of Definition 4.10 can be obtained using additional $O(\kappa\log(1/\epsilon))$ gradients or $O(\epsilon^{-2})$ stochastic gradients. Conversely, if a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ/κ-stationary point in terms of Definition 4.10, a point $\hat{\mathbf{x}}$ is an $O(\epsilon)$-stationary point in terms of Definition 3.3.*

Proposition 4.12 *Under Assumption 4.6, if a point $\hat{\mathbf{x}}$ is an ϵ-stationary point in terms of Definition 3.7, an $O(\epsilon)$-stationary point $(\mathbf{x}', \mathbf{y}')$ in terms of Definition 4.10 can be obtained using additional $O(\epsilon^{-2})$ gradients or $O(\epsilon^{-4})$*

stochastic gradients. Conversely, if a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ^2-stationary point in terms of Definition 4.10, a point $\hat{\mathbf{x}}$ is an $O(\epsilon)$-stationary point in terms of Definition 3.3.

To translate the notion of stationarity based on ∇f to our notion of stationarity, we need to pay an additional factor of $O(\kappa \log(1/\epsilon))$ or $O(\epsilon^{-2})$ in the two settings. In this sense, our notion of stationarity is stronger than the notion based on ∇f in the literature (Lu et al., 2019; Nouiehed et al., 2019). We defer the proofs of these propositions to Appendix B.

4.4. Discussions

Note that the focus of this paper is to provide basic nonasymptotic guarantees for the simple, and widely-used, two-time-scale GDA and SGDA algorithms in the nonconvex-(strongly)-concave settings. We do not wish to imply that these algorithms are optimal in any sense, nor that acceleration should necessarily be achieved by incorporating momentum into the update for the variable $\mathbf{y}$. In fact, the optimal rate for optimizing a nonconvex-(strongly)-concave function remains open. The best known complexity bound has been presented by Thekumparampil et al. (2019) and Kong & Monteiro (2019). Both of the analyses only require $\tilde{O}(\epsilon^{-3})$ gradient computations for solving nonconvex-concave problems but suffer from rather complicated algorithmic schemes. The general question of the construction of optimal algorithms in nonconvex-concave problems is beyond the scope of this paper.

Second, our complexity results are also valid in the convex-concave setting and this does not contradict results showing the divergence of GDA with fixed stepsize. We note a few distinctions: (1) our results guarantee that GDA will visit ϵ-stationary points at some iterates, which are not necessarily the last iterates; (2) our results only guarantee stationarity in terms of $\mathbf{x}_t$, not $(\mathbf{x}_t, \mathbf{y}_t)$. In fact, our proof permits the possibility of significant changes in $\mathbf{y}_t$ even when $\mathbf{x}_t$ is already close to stationarity. This together with our choice $\eta_\mathbf{x} \ll \eta_\mathbf{y}$, makes our results valid. To this end, we highlight that our algorithms can be used to achieve an approximate Nash equilibrium for convex-concave functions (i.e., optimality for both $\mathbf{x}$ and $\mathbf{y}$). Instead of averaging, we run two passes of two-time-scale GDA or SGDA for min-max problem and max-min problem separately. That is, in the first pass we use $\eta_\mathbf{x} \ll \eta_\mathbf{y}$ while in the second pass we use $\eta_\mathbf{x} \gg \eta_\mathbf{y}$. Either pass will return an approximate stationary point for each players, which jointly forms an approximate Nash equilibrium.

5. Overview of Proofs

In this section, we sketch the complexity analysis for two-time-scale GDA (Theorems 4.4 and 4.8).

5.1. Nonconvex-strongly-concave minimax problems

In the nonconvex-strongly-concave setting, our proof involves setting a pair of stepsizes, $(\eta_\mathbf{x}, \eta_\mathbf{y})$, which force $\{\mathbf{x}_t\}_{t \geq 1}$ to move much more slowly than $\{\mathbf{y}_t\}_{t \geq 1}$. Recall Lemma 4.3, which guarantees that $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz:

$$\|\mathbf{y}^\star(\mathbf{x}_1) - \mathbf{y}^\star(\mathbf{x}_2)\| \ \leq \ \kappa\|\mathbf{x}_1 - \mathbf{x}_2\|.$$

If $\{\mathbf{x}_t\}_{t \geq 1}$ moves slowly, then $\{\mathbf{y}^\star(\mathbf{x}_t)\}_{t \geq 1}$ also moves slowly. This allows us to perform gradient ascent on a slowly changing strongly-concave function $f(\mathbf{x}_t, \cdot)$, guaranteeing that $\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_t)\|$ is small in an amortized sense. More precisely, letting the error be $\delta_t = \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2$, the standard analysis of inexact nonconvex gradient descent implies a descent inequality in which the sum of δ_t provides control:

$$\begin{aligned} & \Phi(\mathbf{x}_{T+1}) - \Phi(\mathbf{x}_0) \\ \leq \quad & -\Omega(\eta_\mathbf{x}) \left(\sum_{t=0}^{T} \|\nabla \Phi(\mathbf{x}_t)\|^2\right) + O(\eta_\mathbf{x}\ell^2) \left(\sum_{t=0}^{T} \delta_t\right). \end{aligned}$$

The remaining step is to show that the second term is always small compared to the first term on the right-hand side. This can be done via a recursion for δ_t as follows:

$$\delta_t \ \leq \ \gamma\delta_{t-1} + \beta\|\nabla\Phi(\mathbf{x}_{t-1})\|^2,$$

where $\gamma < 1$ and β is small. Thus, δ_t exhibits a linear contraction and $\sum_{t=0}^{T} \delta_t$ can be controlled by the term $\sum_{t=0}^{T} \|\nabla\Phi(\mathbf{x}_t)\|^2$.

5.2. Nonconvex-concave minimax problems

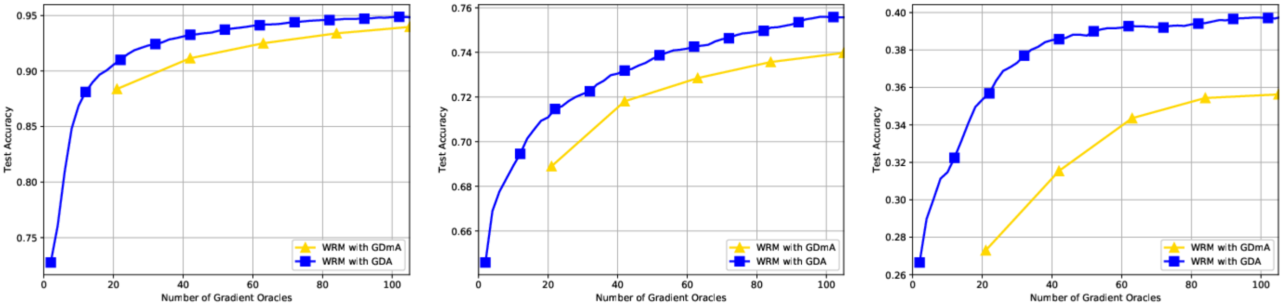
In this setting, the main idea is again to set a pair of learning rates $(\eta_\mathbf{x}, \eta_\mathbf{y})$ which force $\{\mathbf{x}_t\}_{t \geq 1}$ to move more slowly than $\{\mathbf{y}_t\}_{t \geq 1}$. However, $f(\mathbf{x}, \cdot)$ is merely concave and $\mathbf{y}^\star(\cdot)$ is not unique. This means that, even if $\mathbf{x}_1, \mathbf{x}_2$ are extremely close, $\mathbf{y}^\star(\mathbf{x}_1)$ can be dramatically different from $\mathbf{y}^\star(\mathbf{x}_2)$. Thus, $\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_t)\|$ is no longer a viable error to control.

Fortunately, Lemma 4.7 implies that Φ is Lipschitz. That is to say, when the stepsize $\eta_\mathbf{x}$ is very small, $\{\Phi(\mathbf{x}_t)\}_{t \geq 1}$ moves slowly:

$$|\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t-1})| \ \leq \ L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \ \leq \ \eta_\mathbf{x} L^2.$$

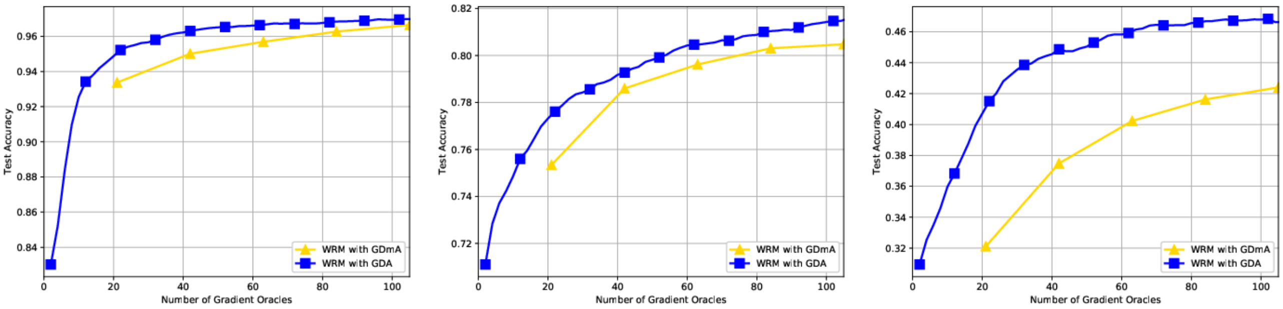
Again, this allows us to perform gradient ascent on a slowly changing concave function $f(\mathbf{x}_t, \cdot)$, and guarantees that $\Delta_t = f(\mathbf{x}_t, \mathbf{z}) - f(\mathbf{x}_t, \mathbf{y}_t)$ is small in an amortized sense where $\mathbf{z} \in \mathbf{y}^\star(\mathbf{x}_t)$. The analysis of inexact nonconvex subgradient descent (Davis & Drusvyatskiy, 2019) implies that Δ_t comes into the following descent inequality:

$$\begin{aligned} \Phi_{1/2\ell}(\mathbf{x}_{T+1}) - \Phi_{1/2\ell}(\mathbf{x}_0) \quad \leq \quad & O(\eta_\mathbf{x}\ell) \left(\sum_{t=0}^{T} \Delta_t\right) \\ & + O(\eta_\mathbf{x}^2 \ell L^2 (T+1)) - O(\eta_\mathbf{x}) \left(\sum_{t=0}^{T} \|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\|^2\right), \end{aligned}$$



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 1. Performance of WRM with GDmA and GDA on MNIST, Fashion-MNIST and CIFAR-10 datasets. We demonstrate test classification accuracy vs. time for different WRM models with GDmA and GDA. Note that $\gamma = 0.4$.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Figure 2. Performance of WRM with GDmA and GDA on MNIST, Fashion-MNIST and CIFAR-10 datasets. We demonstrate test classification accuracy vs. time for different WRM models with GDmA and GDA. Note that $\gamma = 1.3$.

where the first term on the right-hand side is the error term. The remaining step is again to show the error term is small compared to the sum of the first two terms on the right-hand side. To bound the term $\sum_{t=0}^{T} \Delta_t$, we recall the following inequalities and use a telescoping argument (where the optimal point $\mathbf{y}^\star$ does not change):

$$\Delta_t \leq \frac{\|\mathbf{y}_t - \mathbf{y}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{y}^\star\|^2}{\eta_{\mathbf{y}}}. \tag{2}$$

The major challenge here is that the optimal solution $\mathbf{y}^\star(\mathbf{x}_t)$ can change dramatically and the telescoping argument does not go through. An important observation is, however, that (2) can be proved if we replace the $\mathbf{y}^\star$ by any $\mathbf{y} \in \mathcal{Y}$, while paying an additional cost that depends on the difference in function value between $\mathbf{y}^\star$ and $\mathbf{y}$. More specifically, we pick a block of size $B = O(\epsilon^2/\eta_{\mathbf{x}})$ and show that the following statement holds for any $s \leq \forall t < s + B$,

$$\begin{aligned}\Delta_{t-1} \quad &\leq \quad O(\ell)\left(\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right) \\ &\quad +O(\eta_{\mathbf{x}} L^2)(t - 1 - s).\end{aligned}$$

We perform an analysis on the blocks where the concave problems are similar so the telescoping argument can now work. By carefully choosing $\eta_{\mathbf{x}}$, the term $\sum_{t=0}^{T} \Delta_t$ can also be well controlled.

6. Experiments

In this section, we present several empirical results to show that two-time-scale GDA outperforms GDmax. The task is to train the empirical Wasserstein robustness model (WRM) (Sinha et al., 2018) over a collection of data samples $\{\xi_i\}_{i=1}^N$ with ℓ_2-norm attack and a penalty parameter $\gamma > 0$. Formally, we have

$$\min_{\mathbf{x}} \max_{\{\mathbf{y}_i\}_{i=1}^N \subseteq \mathcal{Y}} \frac{1}{N}\left[\sum_{i=1}^{N}\left(\ell(\mathbf{x}, \mathbf{y}_i) - \gamma\|\mathbf{y}_i - \xi_i\|^2\right)\right]. \tag{3}$$

As demonstrated in Sinha et al. (2018), we often choose $\gamma > 0$ sufficiently large such that $\ell(\mathbf{x}, \mathbf{y}_i) - \gamma\|\mathbf{y}_i - \xi_i\|^2$ is strongly concave. To this end, problem (3) is a nonconvex-strongly-concave minimax problem.

We mainly follow the setting of Sinha et al. (2018) and consider training a neural network classifier on three datasets[1]:

[1] https://keras.io/datasets/

MNIST, Fashion-MNIST, and CIFAR-10, with the default cross validation. The architecture consists of 8×8, 6×6 and 5×5 convolutional filter layers with ELU activations followed by a fully connected layer and softmax output. Small and large adversarial perturbation is set with $\gamma \in \{0.4, 1.3\}$ as the same as Sinha et al. (2018). The baseline approach is denoted as *GDmA* in which $\eta_{\mathbf{x}} = \eta_{\mathbf{y}} = 10^{-3}$ and each inner loop contains 20 gradient ascent. Two-time-scale GDA is denoted as *GDA* in which $\eta_{\mathbf{x}} = 5\times 10^{-5}$ and $\eta_{\mathbf{y}} = 10^{-3}$. Figure 1 and 2 show that GDA consistently outperforms GDmA on all datasets. Compared to MNIST and Fashion-MNIST, the improvement on CIFAR-10 is more significant which is worthy further exploration in the future.

7. Conclusion

In this paper, we have shown that two-time-scale GDA and SGDA return an ϵ-stationary point in $O(\kappa^2\epsilon^{-2})$ gradient evaluations and $O(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations in the nonconvex-strongly-concave case, and $O(\epsilon^{-6})$ gradient evaluations and $O(\epsilon^{-8})$ stochastic gradient evaluations in the nonconvex-concave case. Thus, these two algorithms are provably efficient in these settings. In future work we aim to derive a lower bound for the complexity first-order algorithms in nonconvex-concave minimax problems.

Acknowledgements

We would like to thank the area chair and three anonymous referees for constructive suggestions that improve the paper. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In *NeurIPS*, pp. 1576–1584, 2015.

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. Local saddle point optimization: A curvature exploitation approach. *ArXiv Preprint: 1805.05751*, 2018.

Azizian, W., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. A tight and unified analysis of extragradient for a whole spectrum of differentiable games. *ArXiv Preprint: 1906.05945*, 2019.

Bailey, J. P. and Piliouras, G. Multiplicative weights update in zero-sum games. In *EC*, pp. 321–338, 2018.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. *ArXiv Preprint: 1802.05642*, 2018.

Basar, T. and Olsder, G. J. *Dynamic Noncooperative Game Theory*, volume 23. SIAM, 1999.

Benaım, M. and Hirsch, M. W. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1-2):36–72, 1999.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Chen, G. H. G. and Rockafellar, R. T. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Cherukuri, A., Gharesifard, B., and Cortes, J. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55 (1):486–511, 2017.

Daskalakis, C. and Panageas, I. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *ArXiv Preprint: 1807.04252*, 2018a.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *NeurIPS*, pp. 9236–9246, 2018b.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *ArXiv Preprint: 1711.00141*, 2017.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Drusvyatskiy, D. and Lewis, A. S. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Du, S. S. and Hu, W. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. *ArXiv Preprint: 1802.01504*, 2018.

Golshtein, E. G. Generalized gradient method for finding saddle points. *Matekon*, 10(3):36–52, 1974.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, pp. 2672–2680, 2014.

Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. An online learning approach to generative adversarial networks. In *ICLR*, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pp. 6626–6637, 2017.

Hommes, C. H. and Ochea, M. I. Multiple equilibria and limit cycles in evolutionary games with logit dynamics. *Games and Economic Behavior*, 74(1):434–441, 2012.

Jin, C., Netrapalli, P., and Jordan, M. I. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *ArXiv Preprint: 1902.00618*, 2019.

Jordan, M. I. Artificial intelligence–the revolution hasnt happened yet. *Medium. Vgl. Ders.(2018): Perspectives and Challenges. Presentation SysML*, 2018.

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

Kong, W. and Monteiro, R. D. C. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *ArXiv Preprint:1905.13433*, 2019.

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

Kose, T. Solutions of saddle value problems by differential equations. *Econometrica, Journal of the Econometric Society*, pp. 59–70, 1956.

Liang, T. and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *ArXiv Preprint: 1802.06132*, 2018.

Lin, Q., Liu, M., Rafique, H., and Yang, T. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *ArXiv Preprint: 1810.10207*, 2018.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *ArXiv Preprint: 1902.08294*, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ArXiv Preprint: 1706.06083*, 2017.

Mateos, G., Bazerque, J. A., and Giannakis, G. B. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276, 2010.

Mazumdar, E. V., Jordan, M. I., and Sastry, S. S. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *ArXiv Preprint: 1901.00838*, 2019.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Cycles in adversarial regularized learning. In *SODA*, pp. 2703–2717. SIAM, 2018.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. Proximal point approximations achieving a convergence rate of $o(1/k)$ for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *ArXiv Preprint: 1906.01115*, 2019a.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *ArXiv Preprint: 1901.08511*, 2019b.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, pp. 2208–2216, 2016.

Nedić, A. and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

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

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Neumann, J. V. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. *Algorithmic Game Theory*. Cambridge University Press, 2007.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, pp. 14905–14916, 2019.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *ArXiv Preprint: 1810.02060*, 2018.

Robinson, J. An iterative method of solving a game. *Annals of Mathematics*, pp. 296–301, 1951.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, 2015.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training gans with regularized optimal transport. In *NeurIPS*, pp. 7091–7101, 2018.

Shamma, J. *Cooperative Control of Distributed Multi-agent Systems*. John Wiley & Sons, 2008.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pp. 12659–12670, 2019.

Uzawa, H. Iterative methods for concave programming. *Studies in Linear and Nonlinear Programming*, 6:154–165, 1958.

Von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior (Commemorative Edition)*. Princeton University Press, 2007.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510, 2009.