Dissipative Gradient Descent Ascent Method: A Control Theory Inspired Algorithm for Min-max Optimization

Tianqi Zheng, Nicolas Loizou, Pengcheng You, and Enrique Mallada

***Abstract*— Gradient Descent Ascent (GDA) methods for min-max optimization problems typically produce oscillatory behavior that can lead to instability, e.g., in bilinear settings. To address this problem, we introduce a dissipation term into the GDA updates to dampen these oscillations. The proposed Dissipative GDA (DGDA) method can be seen as performing standard GDA on a state-augmented and regularized saddle function that does not strictly introduce additional convexity/concavity. We theoretically show the linear convergence of DGDA in the bilinear and strongly convex-strongly concave settings and assess its performance by comparing DGDA with other methods such as GDA, Extra-Gradient (EG), and Optimistic GDA. Our findings demonstrate that DGDA surpasses these methods, achieving superior convergence rates. We support our claims with two numerical examples that showcase DGDA's effectiveness in solving saddle point problems.**

***Index Terms*— Optimization; Optimization algorithms; Lyapunov methods**

I. Introduction

In recent years, there has been a significant focus on solving saddle point problems, namely min-max optimization problems [1]–[5]. These problems have garnered considerable attention, particularly in fields such as Generative Adversarial Networks (GANs) [5]–[7], Reinforcement Learning (RL) [8], and Constrained RL (C-RL) [9], [10]. However, a major challenge in these approaches is the instability of the training process. That is, solving the min-max optimization problem via running the standard Gradient Descent Ascent (GDA) algorithm often leads to unstable oscillatory behavior rather than convergence to the optimal solution. This is particularly illustrated in bilinear min-max problems, such as the training of Wasserstein GANs [11] or solving C-RL problems in the occupancy measure space [12], for which the standard GDA fails to converge [1], [2].

In order to understand the instability of the GDA method and further tackle its limitation, we draw inspiration from the control-theoretic notions of dissipativity [13], which enables the design of stabilizing controllers using dynamic (state-augmented) components that seek to dissipate the energy generated by the unstable process. This aligns with recent work that leverages control theory tools in the analysis and design of optimization algorithms [14]–[19]. From a dynamical system point of view, dissipativity theory characterizes how energy dissipates within the system and drives it towards equilibrium. It provides a direct way to construct a Lyapunov function, which further relates the rate of decrease of this internal energy to the rate of convergence of the algorithm.

We motivate our developments by looking first at a simple scalar bilinear problem wherein the system's energy, expressed as the square 2-norm distance to the saddle, strictly increases on every iteration, leading to oscillations of increasing amplitude. To tackle this unstable oscillating behavior, we propose the Dissipative GDA method, which, as the name suggests, incorporates a simple friction term to GDA updates to dissipate the internal energy and stabilize the system. Our algorithm can be seen as a discrete-time version of [20], which has been applied to solve the C-RL problems [9]. In this work, we build on this literature, making the following contributions:

1. Novel control theory inspired algorithm: We illustrate how to use control theoretic concepts of dissipativity theory to design an algorithm that can stabilize the unstable behavior of GDA. Particularly, we show that by introducing a friction term, the proposed DGDA algorithm dissipates the stored internal energy and converges toward equilibrium.

2. Theoretical analysis with better rates: We establish the linear convergence of the DGDA method for bilinear and strongly convex-strongly concave saddle point problems. In both settings, we show that the DGDA method outperforms other state-of-the-art first-order explicit methods, surpassing standard known linear convergence rates (see Table I and II).

3. Numerical Validation: We corroborate our theoretical results with numerical experiments by evaluating the performance of the DGDA method with GDA, EG, and OGDA methods. When applied to solve bilinear and strongly convex-strongly concave saddle point problems, the DGDA method systematically outperforms other methods regarding convergence rate.

Outline: The rest of the paper is organized as follows. In Section II, we provide some preliminary definitions and background. In Section III, we leverage tools from dissipativity theory and propose the Dissipative GDA (DGDA) algorithm to tackle the unstable oscillatory behavior of GDA methods. In Section IV, we establish its linear convergence rate for bilinear

T. Zheng, and E. Mallada are with the Department of Electrical and Computer Engineering, N. Loizou is with the Department of Applied Mathematics and Statistics at Johns Hopkins University, Baltimore, MD 21218, USA {tzheng8,nloizou,mallada}@jhu.edu

P. You is with the Department of Industrial Engineering and Management at Peking University, Beijing, China pcyou@pku.edu.cn

This work was supported by NSFC through grants 72201007, 723B1001, T2121002, 72131001, by NSF through grant Global Center 90107717, and Johns Hopkins University IAA Grant Challenge.

TABLE I

GLOBAL CONVERGENCE RESULTS FOR BILINEAR OBJECTIVE FUNCTIONS.

Bilinear	Mokhtari, 20	Azizian, 20	This Work
EG	$\frac{\kappa^{-1}}{20}$	$\frac{\kappa^{-1}}{64}$	-
OG	$\frac{\kappa^{-1}}{800}$	$\frac{\kappa^{-1}}{128}$	-
DG	-	-	$\frac{\kappa^{-1}}{4}$

Summary of the global convergence results for EG, OGDA, and DGDA methods with bilinear objective functions. If a result shows that the iterates converge as $\mathcal{O}((1-r)^t)$, the quantity r is reported (the larger the better). κ represents the condition number.

TABLE II

GLOBAL CONVERGENCE RESULTS FOR STRONGLY CONVEX-STRONGLY CONCAVE AND L-LIPSCHITZ OBJECTIVE FUNCTIONS.

S.C	Zhang, 21	Mokhtari, 20	Azizian, 20	This Work
GD	κ^{-2}	-	-	-
EG	-	$\frac{\kappa^{-1}}{4}$	$\frac{\kappa^{-1}}{4}+\epsilon$	-
OG	-	$\frac{\kappa^{-1}}{4}$	$\frac{\kappa^{-1}}{4}+\epsilon$	-
DG	-	-	-	$\kappa^{-1}-\ \mathcal{O}(\kappa^{-2})$

Summary of the global convergence results for GDA, EG, OGDA, and DGDA methods with strongly convex-strongly concave and L-Lipschitz objective functions. The table reports the term r of a $(1-r)$ linear rate. The constant $\epsilon > 0$ depends on the problem.

and strongly convex-strongly concave problems, which outperforms state-of-the-art first-order explicit algorithms, including GDA, EG, and OGDA methods. In Section V, we support our claims with two numerical examples. We close the paper with concluding remarks and future research directions in Section VI.

II. Problem Formulation

In this paper, we study the problem of finding saddle points in the min-max optimization problem:

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

where the function $f:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is a convex-concave function. Precisely, $f(\cdot,y)$ is convex for all $y\in\mathbb{R}^m$ and $f(x,\cdot)$ is concave for all $x\in\mathbb{R}^n$. We seek to develop a novel optimization algorithm that converges to some saddle point (x^*,y^*) of Problem 1.

Definition 1 (Saddle Point): A point $(x^*,y^*)\in\mathbb{R}^n\times\mathbb{R}^m$ is a saddle point of convex-concave function (1) if and only if it satisfies $f(x^*,y)\le f(x^*,y^*)\le f(x,y^*)$ for all $x\in\mathbb{R}^n, y\in\mathbb{R}^m$.

Throughout this paper, we consider two specific instances of Problem 1 commonly studied in related literature: strongly convex-strongly concave and bilinear functions. Herein, we briefly present some definitions and properties.

Definition 2 (Strongly Convex): A differentiable function $f:\mathbb{R}^n\to\mathbb{R}$ is said to be μ-strongly convex if $f(w)\ge f(w')+\nabla f(w)^T(w-w')+\frac{\mu}{2}\|w-w'\|^2$.

Notice that if $\mu=0$, then we recover the definition of convexity for a continuously differentiable function and $f(w)$ is μ-strongly concave if $-f(w)$ is μ-strongly convex. Another important property commonly used in the convergence analysis of optimization algorithms is the Lipschitz-ness of the gradient $\nabla f(w)$.

Definition 3 (L-Lipschitz): A function $F:\mathbb{R}^n\to\mathbb{R}^m$ is L-Lipschitz if $\forall w,w'\in\mathbb{R}^n$, we have $\|F(w)-F(w')\|\le L\|w-w'\|$.

Combining the above two properties leads to the first important class of problem that has been extensively studied [1], [2], [21], [22].

Assumption 1: (Strongly strongly convex-concave functions with L-Lipschitz Gradient) The function $f:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is continuously differentiable, μ strongly convex in x, and μ strongly concave in y. Further, the gradient vector $(\nabla_x f(x,y);-\nabla_y f(x,y))$ is L-Lipschitz.

It is also crucial to consider situations where the objective function is bilinear. Such bilinear min-max problems often appear when solving constrained reinforcement learning problems [9], [23], and training of WGANs [11].

Assumption 2 (Bilinear function): The function $f:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is a bilinear function if it can be written in the form $f(x,y)=x^TAy$. For simplicity, we further assume that the matrix $A\in\mathbb{R}^{m\times n}$ is full rank, with $m\le n$.

As seen in Table I and II as well as in Section IV, the linear convergence rates of existing algorithms are frequently characterized by the *condition number* κ. Specifically, when the objective function is bilinear, the condition number is defined as $\kappa:=\sigma^2_{\max}(A)/\sigma^2_{\min}(A)$, where $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ denote the largest singular value and smallest singular of a matrix M respectively. When the objective function is strongly convex-strongly concave with the L-Lipschitz gradient, the condition number of the problem is defined as $\kappa:=L/\mu$.

III. Dissipative Gradient Descent Ascent Algorithm

This section introduces the proposed first-order method for solving the min-max optimization problem 1. The algorithm can be seen as a discretization of the algorithm proposed by [20], wherein a regularization framework was introduced for continuous saddle flow dynamics that guarantees asymptotic convergence to a saddle point under mild assumptions. However, the analysis presented in [20] does not generally extend to discrete time. In this paper, we show the linear convergence of the discrete-time version of this algorithm.

Our results build on gaining an intuitive understanding of the problems that one encounters when applying the vanilla GDA method to solve saddle point problems (1):

Gradient Descent Ascent (GDA)

$$x_{k+1}=x_k-\eta\nabla_x f(x_k,y_k),\quad y_{k+1}=y_k+\eta\nabla_y f(x_k,y_k). \tag{2}$$

When (1) is strongly convex-strongly concave with L-Lipschitz gradients, the GDA method provides linear convergence, with step size $\eta=\mu/L^2$ and a know rate estimate of $1-1/\kappa^2$ [24]. However, when the problem is bilinear, the GDA method fails to converge, illustrated in Figure 1.

Our proposed algorithm draws inspiration from dissipative theory in control by introducing two dynamic feedback controllers (friction) to dissipate the energy stored and amplified

by the GDA algorithm. This is implemented in the form of high pass filters of the form

$$\zeta_{k+1} = \zeta_k - \rho(\zeta_k - v_k), \quad w_k = \rho(v_k - \zeta_k), \tag{3}$$

with transfer function $\hat{w}(z) = \frac{z-1}{z-(1-\rho)}\hat{v}(z)$, that is interconnected in negative feedback to attenuate dampen the oscillations of both x_k and y_k. This modification leads to the following proposed algorithm, effectively dampening the oscillations in our illustrative example in Figure 1.

Dissipative gradient descent ascent (DGDA):

$$\begin{bmatrix} x_{k+1} \\ \hat{x}_{k+1} \\ y_{k+1} \\ \hat{y}_{k+1} \end{bmatrix} = \begin{bmatrix} x_k - \eta\nabla_x f(x_k, y_k) - \rho(x_k - \hat{x}_k) \\ \hat{x}_k - \rho(\hat{x}_k - x_k) \\ y_k + \eta\nabla_y f(x_k, y_k) - \rho(y_k - \hat{y}_k) \\ \hat{y}_k - \rho(\hat{y}_k - y_k) \end{bmatrix} \tag{4}$$

Particularly, for f as in (1), in (4) we introduce two new sets of variables $\hat{x} \in \mathbb{R}^n$ and $\hat{y} \in \mathbb{R}^m$ and a damping parameter $\rho > 0$. One important observation is that, due to the high-pass filter structure of the feedback, once the system reaches equilibrium, i.e., $x_{k+1} = x_k$, $y_{k+1} = y_k$, $\hat{x}_{k+1} = \hat{x}_k$, $\hat{y}_{k+1} = \hat{y}_k$, one necessarily has $\hat{x}_k = x_k$ and $\hat{y}_k = y_k$, which ensures that the fixed point is necessarily a critical point of the saddle function.

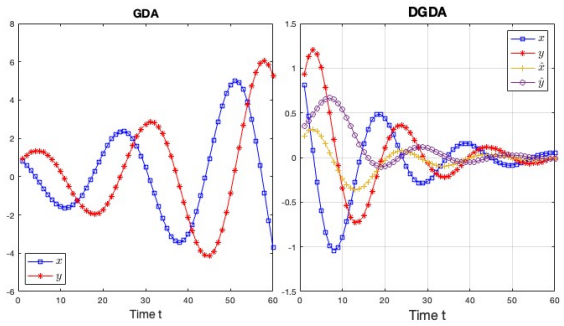

Fig. 1. Trajectories of states for GDA and DGDA for the simple bilinear objective function $f(x, y) := xy$.

The first important observation is that the above DGDA update could be considered as applying a vanilla GDA update to the following regularized surrogate for $f(x, y)$:

$$f(x, y, \hat{x}, \hat{y}) := f(x, y) + \frac{\rho}{2}\|x - \hat{x}\|^2 - \frac{\rho}{2}\|y - \hat{y}\|^2. \tag{5}$$

While our algorithm also introduces two regularizing terms, the following Lemma verifies the fixed positions of saddle points between $f(x, y)$ and $f(x, y, \hat{x}, \hat{y})$ with virtual variables aligned with original variables.

Lemma 1 (Saddle Point Invariance): [20, Lemma 6] For problem 1, a point (x^*, y^*) is a saddle point of $f(x, y)$ if and only if $(x^*, y^*, \hat{x}^*, \hat{y}^*)$ is a saddle point of $f(x, y, \hat{x}, \hat{y})$, with $\hat{x}^* = x^*$ and $\hat{y}^* = y^*$.

More interestingly, the regularization terms, $\frac{\rho}{2}\|x - \hat{x}\|^2$ and $\frac{\rho}{2}\|y - \hat{y}\|^2$, do not introduce extra strong convexity-stong concavity to the original problem. Precisely, the augmented problem $f(x, y, \hat{x}, \hat{y})$ is neither strongly convex on $(x, \hat{x})$ nor strongly concave on $(y, \hat{y})$. Indeed, on the hyperplane of $x = \hat{x}$ and $y = \hat{y}$, the augmented problem recovers the original problem $f(x, y, \hat{x}, \hat{y}) = f(x, y)$.

We finalize this section by comparing DGDA with recent efforts to solve min-max optimization problems. We note that DGDA is different from the *Proximal Point Method* [21] or introducing a L_2 *regularization* [25]. Notably, in [26] they introduce an accelerated proximal point method, MINIMAX-APPA that has $\tilde{\mathcal{O}}(\sqrt{\kappa_x\kappa_y})$ gradient complexity, matching the theoretical lower bound up to logarithmic factors. In the following section, we will show that our proposed algorithm gets a comparable and slightly better complexity bound $\mathcal{O}(\sqrt{\kappa_x\kappa_y})$, while we do not require x, y to belongs to bounded sets.

Recent research has also utilized *Moreau-Yosida* smoothing techniques to tackle various optimization problems, ranging from nonconvex-concave [27]–[29] to nonconvex-nonconcave optimization problems [30]. These approaches also fall under the category of first-order *Implicit* methods. In this work, we focus on comparing with first-order *Explicit* algorithms. While our primary focus lies on strongly convex, strongly concave, and bilinear settings, we also delve into further analyses across other contexts, including nonconvex [27]–[30] and stochastic settings [31], [32].

IV. Convergence Analysis

In this section, we provide a theoretical analysis of the proposed algorithm. Consider a quadratic Lyapunov function to track the energy dissipation of the DGDA updates

$$V_k := \|x_k - x^*\|^2 + \|y_k - y^*\|^2 + \|\hat{x}_k - \hat{x}^*\|^2 + \|\hat{y}_k - \hat{y}^*\|^2,$$

which denotes the square 2-norm distance to the saddle point at the k-th iteration. The goal is, therefore, to find some $0 \leq \alpha < 1$ such that $V_{k+1} \leq \alpha V_k$, where α denotes the linear convergence rate.

A. Convergence Analysis for Bilinear Functions

When applied to the bilinear min-max optimization problem $f(x, y) = x^T Ay$, the DGDA update (4) is equivalent to a linear dynamical system. Specifically, denote $z = [x, y]^T, \hat{z} = [\hat{x}, \hat{y}]^T$ yields:

$$\begin{bmatrix} z_{k+1} - z^* \\ \hat{z}_{k+1} - \hat{z}^* \end{bmatrix} = \begin{bmatrix} (1-\rho)I - \eta M & \rho I \\ \rho I & (1-\rho)I \end{bmatrix} \begin{bmatrix} z_k - z^* \\ \hat{z}_k - \hat{z}^* \end{bmatrix}, \tag{6}$$

where $M = \begin{bmatrix} \mathbf{0} & A \\ -A^T & \mathbf{0} \end{bmatrix}$. Therefore, the linear convergence rate of DGDA can be derived from the analysis of the spectrum of the associated matrix that defines the DGDA update in (6). This yields the following theorem.

Theorem 2: (Linear convergence of DGDA, Bilinear Case) Let Assumption 2 hold. Then the updates 4 of DGDA with $0 < \eta \leq \frac{2\rho}{\sigma_{\max}(A)}$ and $\rho > 0$ provide linearly converging iterates:

$$V_k \leq \mathcal{O}\left(\left(1 - 2\rho + 2\rho^2 + (1-\rho)\sqrt{4\rho^2 - \eta^2\sigma_{\min}^2(A)}\right)^k\right) V_0,$$

Particularly, setting $\rho = 1/2$ and $\eta = 1/\sigma_{\max}(A)$ we have

$$V_k \leq \mathcal{O}\left(\left(1 - \tfrac{1}{4\kappa}\right)^k\right) V_0. \tag{7}$$

Proof: We consider, for ease of presentation, the case when $A \in \mathbb{R}^{m\times m}$ is a square non-singular matrix, i.e., the point $(x^*, y^*) = (\mathbf{0}, \mathbf{0})$ is the unique saddle point. The extension for non-square matrices is straightforward and has been covered in the literature [33, Appendix G]. According to [2, Lemma 7], we have $\mathrm{Sp}(M) = \{\pm i\sigma | \sigma^2 \in \mathrm{Sp}(AA^T)\}$. Therefore, we can compute the eigenvalues of system (6):

$$\mu_j = 1 - \rho \pm i(\frac{1}{2}\eta\sigma_j) \pm \frac{1}{2}\sqrt{4\rho^2 - \eta^2\sigma_j^2}, \tag{8}$$

where $\pm i\sigma_j \in \mathrm{Sp}(M)$. Suppose that for all $j \in [m]$, we choose $0 < \eta \leq \frac{2\rho}{\sigma_{\max}} \leq \frac{2\rho}{\sigma_j}$ and $\rho > 0$, which implies $4\rho^2 - \eta^2\sigma_j^2 \geq 0$, then we can construct the following upper bound for the magnitude of eigenvalues,

$$|\mu_j|^2 = 1 - 2\rho + 2\rho^2 \pm (1-\rho)\sqrt{4\rho^2 - \eta^2\sigma_j^2} \tag{9}$$

$$< 1 - 2\rho + 2\rho^2 + (1-\rho)\sqrt{4\rho^2} \quad = 1 \ . \tag{10}$$

It follows from standard linear systems theory, e.g. [34, Theorem 8.3], the above spectral radius analysis of the linear system (6) results in the following linear convergence rate estimate:

$$V_k \leq \mathcal{O}\left(\left(1 - 2\rho + 2\rho^2 + (1-\rho)\sqrt{4\rho^2 - \eta^2\sigma_{\min}^2}\right)^k\right) V_0,$$

where $V_k := \|x_k - x^*\|^2 + \|y_k - y^*\|^2 + \|\hat{x}_k - \hat{x}^*\|^2 + \|\hat{y}_k - \hat{y}^*\|^2$. Furthermore, the analysis of the above bound identifies the following optimal step sizes $\eta = \frac{2\rho}{\sigma_{\max}}$ and $\rho = \frac{1}{2}$, and the following linear convergence rate estimate

$$V_k \leq \mathcal{O}\left(\left(1 - \tfrac{1}{4\kappa}\right)^k\right) V_0. \tag{11}$$

■

We remark that linear convergence requires $\rho > 0$. This is not surprising since GDA, which is known to diverge for bilinear functions, can be interpreted as the DGDA method when $\rho = 0$. More importantly, by choosing the optimal step size $\rho = 1/2, \eta = 1/\sigma_{\max}(A)$, DGDA method achieves a better linear convergence rate than the EG and OGDA methods (see Table I).

B. Convergence Analysis for Strongly Convex Stronly Concave Functions

We now consider the case of strongly convex-strongly concave min-max problems. Let $F(z_k) := (\nabla_x f(x_k, y_k), -\nabla_y f(x_k, y_k))$. The DGDA updates can be written as follows:

$$\begin{bmatrix} z_{k+1} \\ \hat{z}_{k+1} \end{bmatrix} = \begin{bmatrix} z_k - \eta F(z_k) - \rho(z_k - \hat{z}_k) \\ \hat{z}_k - \rho(\hat{z}_k - z_k) \end{bmatrix} \tag{12}$$

Because of the existence of the nonlinear term $F(z_k)$, we cannot analyze the spectrum as in the previous bilinear case. This is indeed a common challenge in analyzing most optimization algorithms beyond a neighborhood of the fixed point. We circumvent this problem by leveraging recent results on the analysis of variational mappings as $F(\cdot)$ via integral quadratic constraint [15]–[17].

Theorem 3: (Linear convergence of DGDA, Strongly Convex-Strongly Concave Case) Let Assumption 1 hold, then the updates (4) with $\rho = 1/2$ and $\eta = 1/(L+\mu)$ of the DGDA algorithm provide linearly converging iterates:

$$V_k \leq \left(1 - \kappa^{-1} + \mathcal{O}(\kappa^{-2})\right)^k V_0 \tag{13}$$

Proof: Given a linear dynamical system of the form: $\xi_{k+1} = A\xi_k + Bw_k$, where $\xi \in \mathbb{R}^{n_\xi}$ is the state, $w_k \in \mathbb{R}^{n_w}$ is the input, A is the state transition matrix and B is the input matrix. Suppose that there exist a (Lyapunov) function V, satisfying $V(\xi) \geq 0, \forall \xi \in \mathbb{R}^{n_\xi}$, some $0 \leq \alpha < 1$ and a supply rate function $S(\xi_k, w_k) \leq 0, \forall k$ such that

$$V(\xi_{k+1}) - \alpha^2 V(\xi_k) \leq S(\xi_k, w_k), \tag{14}$$

then this dissipation inequality (14) implies that $V(\xi_{k+1}) \leq \alpha^2 V(\xi_k)$, and the state will approach a minimum value at equilibrium no slower than the linear rate α^2 [15]. According to [17, Lemma 6], we could construct the following Linear Matrix Inequality and supply rate function for DGDA updates, by augmenting the states $\xi_k = (z_k; \hat{z}_k)$,

$$S(\xi_k, w_k) = \begin{bmatrix} z_k \\ \hat{z}_k \\ w_k \end{bmatrix}^T \begin{bmatrix} 2\mu L I & 0 & (-\mu + L)I \\ 0 & 0 & 0 \\ (-\mu+L)I & 0 & 2I \end{bmatrix} \begin{bmatrix} z_k \\ \hat{z}_k \\ w_k \end{bmatrix} \leq 0 \tag{15}$$

where the nonlinear operator $F(z_k)$ meets the conditions specified in Assumption 1.

Finally, according to [15, Theorem 2], constructing the dissipation inequality (14) and proving linear convergence can be achieved through solving a semidefinite programming problem. Precisely, if there exists matrix $X^T = X$ and $P \in \mathbb{R}^{n_\xi \times n_\xi}$ with $P \succeq 0$ such that

$$\begin{bmatrix} A^T P A - \alpha^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} - X \leq 0, \tag{16}$$

where $S(\xi, w) := \begin{bmatrix} \xi \\ w \end{bmatrix}^T X \begin{bmatrix} \xi \\ w \end{bmatrix}$, then the dissipation inequality holds for all trajectories of $\xi_{k+1} = A\xi_k + Bw_k$, with $V(\xi) = \xi^T P \xi$. Given the set of problem parameters, a set of feasible solutions is given by:

$$\rho = \frac{1}{2}, \eta = \frac{1}{L+\mu}, P = \begin{bmatrix} (L+\mu)^2 & 0 \\ 0 & (L+\mu)^2 \end{bmatrix} \otimes I, \tag{17}$$

$$\alpha^2 = \frac{3L^2 + 2L\mu + 3\mu^2 + \sqrt{(L+\mu)^4 + 16L^2\mu^2}}{4(L+\mu)^2}. \tag{18}$$

After substituting the condition number $\kappa := L/\mu$, the convergence rate simplifies to $\alpha^2 = 1 - \kappa^{-1} + \mathcal{O}\big((\frac{\mu}{L})^2\big)$ ■

Similarly, as in the bilinear case, we remark on the importance of the dissipation component. When $\rho = 0$, a similar analysis as in the proof of the theorem recovers the lower bound of the convergence rate of GDA $(1 - \kappa^{-2})$ as shown in [18, 3.1]. Thus, our DGDA method provides a better convergence rate estimate than GDA, since clearly $\kappa \in [1, \infty)$, and therefore $\kappa^{-2} \leq \kappa^{-1}$. Additionally, Theorem 2 and Theorem 3 indicate that if we want to achieve an ϵ-accurate solution, we need to run at most $\mathcal{O}(\kappa \log(1/\epsilon))$ iterations (gradient evaluations).

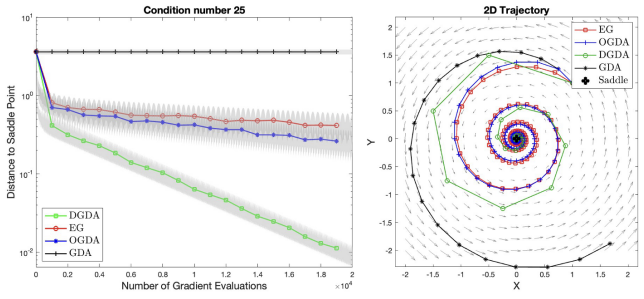

Fig. 2. Convergence of GDA, EG, OGDA, and DGDA in terms of the number of gradient evaluations for the bilinear problem. GDA diverges and the error is not shown. All other three algorithms converge linearly, where the DGDA method provides the best performance.

We remark that while the rate obtained in Theorem 3 is better than those of the EG and OGDA methods for large condition numbers κ (see Table II), the theorem fails to quantify the comparative performance of DGDA for small values of κ. The following corollary shows that indeed, the rate of DGDA is provably better for all $\kappa \geq 2$.

Corollary 4 (SCSC, comparison with known rates): Let Assumption 1 hold, and suppose that $L \geq 2m$, i.e., $\kappa \geq 2$. Then, the linear convergence rate estimate of DGDA (13) is smaller (better) than that of EG and OGDA, i.e., $1 - \kappa^{-1}/4$ (Theorem 6&7 [2] and Theorem 4&7 [1]).

V. Numerical Experiments

In this section, we compare the performance of the proposed Dissipative gradient descent (DGDA) method with the Extra-gradient (EG), Gradient descent ascent (GDA), and Optimistic gradient descent ascent (OGDA) methods.

A. Bilinear problem

We first consider the following bilinear min-max optimization problem: $\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} x^T A y$, where $A \in \mathbb{R}^{m\times n}$ is full-rank. The simulation results are illustrated in Figure 2. In this experiment, we set the dimension of the problem to $m = n = 10$ and the iterates are initialized at x_0, y_0, which are randomly drawn from the uniform distribution on the open interval $(0, 1)$.

We plot the errors (distance to saddle points) of DGDA, EG, and OGDA versus the number of gradient evaluations for this problem in the left plot of Figure 2. The solid line and grey-shaded error bars represent the average trajectories and standard deviations of 20 trials, where in each trial the randomly generated matrix A has a fixed condition number, i.e., $\kappa = \sigma^2_{\max}(A)/\sigma^2_{\min}(A) = 25$. The key motivation is that all three algorithms' convergence rates critically depend on κ^{-1}, and by fixing the condition number, we provide an explicit comparison of their convergence speed.

We pick the step size for different methods according to theoretical findings. That is, we select $\rho = 1/2$ and $\eta = 1/\sigma_{\max}(A)$ for DGDA (Theorem 2), $\eta = 1/4L = 1/4\sigma_{\max}(A)$ for EG and OGDA (Theorem 6&7 [2] and Theorem 4&7 [1]). We do not show the error of GDA since it diverges for this bilinear saddle point problem. All other three algorithms converge linearly, with the DGDA method providing the best performance.

Finally, to provide a qualitative demonstration of how DGDA fares with other existing algorithms, we further plot the sample trajectories of GDA, EG, OGDA, and EGDA on a simple 2D bilinear min-max problem, with $m = n = 1$. In right plot of Figure 2, we observe that while GDA diverges, the trajectories of all other three algorithms converge linearly to the saddle point $(x^*, y^*) = (0, 0)$. Interestingly, our proposed algorithm (DGDA) despite taking larger steps, exhibits faster linear convergence.

B. Strongly convex-strongly concave problem

In the second numerical example, we focus on a strongly convex-strongly strongly-concave quadratic problem of the following form:

$$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} \frac{1}{2}x^T A x - \frac{1}{2}y^T B y + x^T C y, \tag{19}$$

where the matrices satisfy $\mu_A I \preceq A \preceq L_A I$, $\mu_B I \preceq B \preceq L_B I$, $\mu_c^2 I \preceq C^T C \preceq L_c^2 I$. As a result, the problem (19) satisfy Assumption 1. In this experiment, we set the dimension of the problem to $n = 50, m = 10$, and the iterates are initialized at x_0, y_0, which are randomly drawn from the uniform distribution on the open interval $(0, 1)$. We plot the errors (distance to saddle points) of GDA, DGDA, EG, and OGDA versus the number of gradient evaluations for this problem in Figure 3. Again, the solid line and grey-shaded error bars represent the average trajectories and standard deviations of 20 trials, where in each trial the randomly generated matrix $\begin{bmatrix} A & C \\ -C^T & B \end{bmatrix}$ is chosen such that the condition number of (19) remains constant, i.e., $\kappa = L/\mu = 31$. Similarly as in the bilinear problem in Section V-A, we pick the step size for the DGDA method according to our theoretical finding in Theorem 3. The step size of the GDA method is selected as $\eta = \mu/L^2$ (Theorem 5 [35]). The step sizes for EG and OGDA methods are selected as $\eta = 1/4L$ (Theorem 6&7 [2] and Theorem 4&7 [1]). According to the plots, all algorithms converge linearly, and the DGDA method has the best performance.

VI. Conclusion and Future Work

In this work, we present the Dissipative GDA (DGDA) algorithm, a novel method for solving min-max optimization problems. Drawing inspiration from dissipativity theory and control theory, we address the challenge of diverging oscillations in bilinear min-max optimization problems when using the Gradient Descent Ascent (GDA) method. Particularly, we introduce a friction term into the GDA updates aiming to dissipate the internal energy and drive the system towards equilibrium. By incorporating a state-augmented regularization, our proposed DGDA method can be seen as performing standard GDA on an extended saddle function without introducing additional convexity. We further establish the superiority of the convergence rate of the proposed DGDA method when compared with other established methods including GDA,

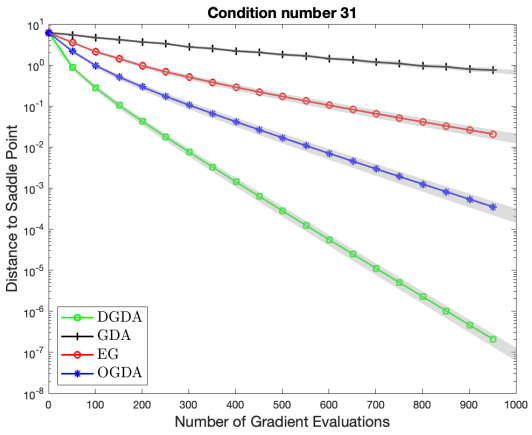

Fig. 3. Convergence of GDA, EG, OGDA, and DGDA in terms of the number of gradient evaluations for problem 19. All algorithms converge linearly, and the DGDA method has the best performance.

Extra-Gradient (EG), and Optimistic GDA. The analysis is further supported by two numerical examples, demonstrating its effectiveness in solving saddle point problems. Our future work includes studying the DGDA method in a stochastic setting and its application in solving Constrained Reinforcement learning problems in the policy space.

REFERENCES

[1] A. Mokhtari, A. Ozdaglar, and S. Pattathil, "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1497–1507.

[2] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel, "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2863–2873.

[3] N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien, "Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 095–19 108, 2021.

[4] A. Beznosikov, E. Gorbunov, H. Berard, and N. Loizou, "Stochastic gradient descent-ascent: Unified theory and new efficient methods," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 172–235.

[5] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien, "A variational inequality perspective on generative adversarial networks," *arXiv preprint arXiv:1802.10551*, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[7] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training gans with optimism," *arXiv preprint arXiv:1711.00141*, 2017.

[8] D. Pfau and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *arXiv preprint arXiv:1610.01945*, 2016.

[9] T. Zheng, P. You, and E. Mallada, "Constrained reinforcement learning via dissipative saddle flow dynamics," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2022, pp. 1362–1366.

[10] D. Ding, C.-Y. Wei, K. Zhang, and A. Ribeiro, "Last-iterate convergent policy gradient primal-dual methods for constrained mdps," *arXiv preprint arXiv:2306.11700*, 2023.

[11] J. Adler and S. Lunz, "Banach wasserstein gan," *Advances in neural information processing systems*, vol. 31, 2018.

[12] E. Altman, *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

[13] S. Sastry, *Nonlinear systems: analysis, stability, and control*. Springer Science & Business Media, 2013, vol. 10.

[14] Z. E. Nelson and E. Mallada, "An integral quadratic constraint framework for real-time steady-state optimization of linear time-invariant systems," in *2018 annual American control conference (ACC)*. IEEE, 2018, pp. 597–603.

[15] B. Hu and L. Lessard, "Dissipativity theory for nesterov's accelerated method," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1549–1557.

[16] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2654–2689, 2018.

[17] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[18] G. Zhang, X. Bao, L. Lessard, and R. Grosse, "A unified analysis of first-order methods for smooth games via integral quadratic constraints," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4648–4686, 2021.

[19] M. Doostmohammadian, W. Jiang, M. Liaquat, A. Aghasi, and H. Zarrabi, "Discretized distributed optimization over dynamic digraphs," *IEEE Transactions on Automation Science and Engineering*, 2024.

[20] P. You and E. Mallada, "Saddle flow dynamics: Observable certificates and separable regularization," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 4817–4823.

[21] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM journal on control and optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[22] P. Tseng, "On linear convergence of iterative methods for the variational inequality problem," *Journal of Computational and Applied Mathematics*, vol. 60, no. 1-2, pp. 237–252, 1995.

[23] M. Wang, "Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time," *Mathematics of Operations Research*, vol. 45, no. 2, pp. 517–546, 2020.

[24] B. Grimmer, H. Lu, P. Worah, and V. Mirrokni, "The landscape of the proximal point method for nonconvex–nonconcave minimax optimization," *Mathematical Programming*, vol. 201, no. 1-2, pp. 373–407, 2023.

[25] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[26] T. Lin, C. Jin, and M. I. Jordan, "Near-optimal algorithms for minimax optimization," in *Conference on Learning Theory*. PMLR, 2020, pp. 2738–2779.

[27] J. Zhang, P. Xiao, R. Sun, and Z. Luo, "A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems," *Advances in neural information processing systems*, vol. 33, pp. 7377–7389, 2020.

[28] Z. Xu, H. Zhang, Y. Xu, and G. Lan, "A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems," *Mathematical Programming*, pp. 1–72, 2023.

[29] J. Yang, A. Orvieto, A. Lucchi, and N. He, "Faster single-loop algorithms for minimax optimization without strong concavity," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 5485–5517.

[30] T. Zheng, L. Zhu, A. M.-C. So, J. Blanchet, and J. Li, "Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[31] W. Xian, F. Huang, Y. Zhang, and H. Huang, "A faster decentralized algorithm for nonconvex minimax problems," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 865–25 877, 2021.

[32] F. Huang, S. Gao, J. Pei, and H. Huang, "Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization," *Journal of Machine Learning Research*, vol. 23, no. 36, pp. 1–70, 2022.

[33] G. Zhang and Y. Yu, "Convergence of gradient methods on bilinear zero-sum games," in *International Conference on Learning Representations*, 2019.

[34] J. P. Hespanha, *Linear systems theory*. Princeton university press, 2018.

[35] A. Beznosikov, B. Polyak, E. Gorbunov, D. Kovalev, and A. Gasnikov, "Smooth monotone stochastic variational inequalities and saddle point problems: A survey," *European Mathematical Society Magazine*, no. 127, pp. 15–28, 2023.