Solving Min-Max Optimization with Hidden Structure via Gradient Descent Ascent

Lampros Flokas*
Department of Computer Science
Columbia University
New York, NY 10025
lamflokas@cs.columbia.edu

Emmanouil V. Vlatakis-Gkaragkounis*
Department of Computer Science
Columbia University
New York, NY 10025
emvlatakis@cs.columbia.edu

Georgios Piliouras
Singapore University of Technology & Design
georgios.piliouras@sutd.edu.sg

Abstract

Many recent AI architectures are inspired by zero-sum games, however, the behavior of their dynamics is still not well understood. Inspired by this, we study standard gradient descent ascent (GDA) dynamics in a specific class of non-convex non-concave zero-sum games, that we call hidden zero-sum games. In this class, players control the inputs of smooth but possibly non-linear functions whose outputs are being applied as inputs to a convex-concave game. Unlike general zero-sum games, these games have a well-defined notion of solution; outcomes that implement the von-Neumann equilibrium of the “hidden" convex-concave game. We provide conditions under which vanilla GDA provably converges not merely to local Nash, but the actual von-Neumann solution. If the hidden game lacks strict convexity properties, GDA may fail to converge to any equilibrium, however, by applying standard regularization techniques we can prove convergence to a von-Neumann solution of a slightly perturbed zero-sum game. Our convergence results are non-local despite working in the setting of non-convex non-concave games. Critically, under proper assumptions we combine the Center-Stable Manifold Theorem along with novel type of initialization dependent Lyapunov functions to prove that almost all initial conditions converge to the solution. Finally, we discuss diverse applications of our framework ranging from generative adversarial networks to evolutionary biology.

1 Introduction

Traditionally, our understanding of convex-concave games revolves around von Neumann’s celebrated minimax theorem, which implies the existence of saddle point solutions with a uniquely defined value. These solutions are called *von Nemann solutions* and guarantee to each agent their corresponding value regardless of opponent play. Although many learning algorithms are known to be able to compute such saddle points [13], recently there has there has been a fervor of activity in proving stronger results such as faster regret minimization rates or analysis of the day-to-day behavior [46, 17, 7, 1, 66, 19, 2, 45, 5, 25, 70, 29, 6, 48, 30, 56].

This interest has been largely triggered by the impressive successes of AI architectures inspired by min-max games such as Generative Adversarial Networks (GANS) [26], adversarial training [40] and reinforcement learning self-play in games [63]. Critically, however, all these applications are based upon *non-convex non-concave games*, our understanding of which is still nascent. Nevertheless,

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

some important early work in the area has focused on identifying new solution concepts that are widely applicable in general min-max games, such as (local/differential) Nash equilibrium [3, 41], local minmax [18], local minimax [31], (local/differential) Stackleberg equilibrium [24], local robust point [69]. The plethora of solutions concepts is perhaps suggestive that "solving" general min-max games unequivocally may be too ambitious a task. Attraction to spurious fixed points [18], cycles [65], robustly chaotic behavior [15, 16] and computational hardness issues [20] all suggest that general min-max games might inherently involve messy, unpredictable and complex behavior.

Are there rich classes of non-convex non-concave games with an effectively unique game theoretic solution that is selected by standard optimization dynamics (e.g. gradient descent)?

Our class of games. We will define a general class of min-max optimization problems, where each agent selects its own vectors of parameters which are then processed separately by smooth functions. Each agent receives their respective payoff after entering the outputs of the processed decision vectors as inputs to a standard convex-concave game. Formally, there exist functions $F : \mathbb{R}^N \to \mathcal{X} \subset \mathbb{R}^n$ and $G : \mathbb{R}^M \to \mathcal{Y} \subset \mathbb{R}^m$ and a continuous convex-concave function $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that the min-max game is

$$\min_{\theta \in \mathbb{R}^N} \max_{\phi \in \mathbb{R}^M} L(F(\theta), G(\phi)). \qquad \text{(Hidden Convex-Concave (HCC))}$$

We call this class of min-max problems Hidden Convex-Concave Games. It generalizes the recently defined hidden bilinear games of [65].

Our solution concept. Out of all the local Nash equilibria of HCC games, there exists a special subclass, the vectors (θ^*, ϕ^*) that implement the von Neumann solution of the convex-concave game. This solution has a strong and intuitive game theoretic justification. Indeed, it is stable even if the agents could perform arbitrary deviations directly on the output spaces $\mathcal{X}, \mathcal{Y}$. These parameter combinations (θ^*, ϕ^*) "solve" the "hidden" convex-concave L and thus we call them *von Neumann solutions*. Naturally, HCCs will typically have numerous local saddle/Nash equilibria/fixed points that do not satisfy this property. Instead, they correspond to stationary points of the F, G where their output is stuck, e.g., due to an unfortunate initialization. At these points the agents may be receiving payoffs which can be arbitrarily smaller/larger than the game theoretic value of game L. Fortunately, we show that Gradient Descent Ascent (GDA) strongly favors von Neumann solutions over generic fixed points.

Our results. In this work, we study the behavior of continuous GDA dynamics for the class of HCC games where each coordinate of F, G is controlled by disjoint sets of variables. In a nutshell, we show that GDA trajectories stabilize around or converge to the corresponding von Neumann solutions of the hidden game. Despite restricting our attention to a subset of HCC games, our analysis has to overcome unique hurdles not shared by standard convex concave games.

Challenges of HCC games. In convex-concave games, deriving the stability of the von Neumann solutions relies on the Euclidean distance from the equilibrium being a Lyapunov function. In contrast, in HCC games where optimization happens in the parameter space θ, ϕ, the non-linear nature of F, G distorts the convex-concave landscape in the output space. Thus, the Euclidean distance will not be in general a Lyapunov function. Moreover, the existence of any Lyapunov function for the trajectories in the output space of F, G does not translate to a well-defined function in the parameter space (unless F, G are trivial, invertible maps). Worse yet, even if L has a unique solution in the output space, this solution could be implemented by multiple equilibria in the parameter space and thus each of them can not be individually globally attracting. Clearly any transfer of stability or convergence properties from the output to the parameter space needs to be initialization dependent. It is worth mentioning that similar challenges like transfering results from the output to the input space was also faced in the simpler class of hidden bilinear games. However, [65] to sidestep this issue assume the restricitve requirement of F, G to be invertible operators. Our results go beyond this simplified case requiring new proof techniques. Specifically, we show how to combine the powerful technologies of the the Center-Stable Manifold Theorem, typically used to argue convergence to equilibria in non-convex optimization settings [34, 52, 54, 53, 35], along with a novel Lyapunov function argument to prove that almost all initial conditions converge to the our game theoretic solution.

Lyapunov Stability. Our first step is to construct an initialization-dependent Lyapunov function that accounts for the curvature induced by the operators F and G (Lemma 2). Leveraging a potentially infinite number of initialization-dependent Lyapunov functions in Theorem 5 we prove that under mild assumptions the outputs of F, G stabilize around the von Neumann solution of L.

Convergence Mirroring convex concave games, we require strict convexity or concavity of L to provide convergence guarantees to von Neumann solutions (Theorem 6). Barring initializations where von Neumann solutions are not reachable due to the limitations imposed by F and G, the set of von Neumann solutions are globally asymptotically stable (Corollary 1). Even in non-strict HCC games, we can add regularization terms to make L strictly convex concave. Small amounts of regularization allows for convergence without significantly perturbing the von Neumann solution (Theorem 7) while increasing regularization enables exponentially faster convergence rates (Theorem 8). Similar to the aforementioned theoretical work, our model of HCC games provides a formal and theoretical tractable testbed for evaluating the performance of different training methods in GAN inspired architectures. As a concrete example, [36] recently proved the success of WGAN training for learning the parameters of non-linearly transformed Gaussian distributions, where for simplicity they replaced the typical Lipschitz constraint of the discriminator function with a quadratic regularizer. Interestingly, we can elucidate on why regularized learning is actually necessary by establishing a formal connection to HCC games. On top of other such ML applications, our game theoretic framework can furthermore capture and generalize evolutionary game theoretic models. [57] analyze a model of evolutionary competition between two species (host-parasite). The outcome of this competition depends on their respective phenotypes (informally their properties, e.g., agility, camouflage, etc.). These phenotypes are encoded via functions that map input vectors (here genotype/DNA sequences) to phenotypes. While [57] proved that learning in these games does not converge to equilibria and typically cycles for almost all initial conditions, we can explicitly construct initial conditions that do not satisfy our definition of safety and end up converging to artificial fixed points. Safety conditions aside, we show that a slight variation of the evolutionary/learning algorithm suffices to resolve the cycling issues and for the dynamics to equilibrate to the von Neumann solution. Hence, we provide the first instance of team zero-sum games [62], a notoriously hard generalization of zero-sum games with a large duality gap, that is solvable by decentralized dynamics.

Organization. In Section 2 we provide some preliminary notation, the definition of our model and some useful technical lemmas. Section 3 is devoted to the presentation of our the main results. Section 4 discusses applications of our framework to specific GAN formulations. Section 5 concludes our work with a discussion of future directions and challenges. We defer the full proofs of our results as well as further discussion on applications to the Appendix.

2 Preliminaries

2.1 Notation

Vectors are denoted in boldface $\mathbf{x}, \mathbf{y}$ unless otherwise indicated are considered as column vectors. We use $\|\cdot\|$ to denote the ℓ_2 norm. For a function $f : \mathbb{R}^d \to \mathbb{R}$ we use ∇f to denote its gradient. For functions of two vector arguments, $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, we use $\nabla_{\mathbf{x}} f, \nabla_{\mathbf{y}} f$ to denote its partial gradient. For the time derivative we will use the dot accent abbreviation, i.e., $\dot{\mathbf{x}} = \frac{d}{dt}[\mathbf{x}(t)]$. A function f will belong to C^r if it is r times continuously differentiable. Additionally, $f \circ g = f(g(\cdot))$ denotes the composition of f, g. Finally, the term "sigmoid" function refers to $\sigma : \mathbb{R} \to \mathbb{R}$ such that $\sigma(x) = (1 + e^{-x})^{-1}$.

2.2 Hidden Convex Concave Games

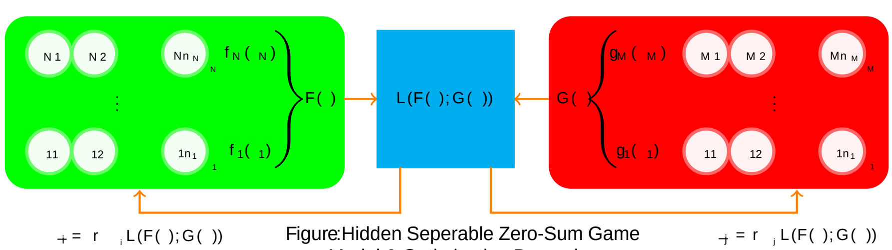

Figure: Hidden Seperable Zero-Sum Game Model & Optimization Dynamics

We will begin our discussion by defining the notion of convex concave functions as well as strictly convex concave functions. Note that our definition of strictly convex concave functions is a superset of strictly convex strictly concave functions that are usually studied in the literature.

Definition 1. $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex concave if for every $y \in \mathbb{R}^n$ $L(\cdot; y)$ is convex and for every $x \in \mathbb{R}^m$ $L(x; \cdot)$ is concave. Function L will be called strictly convex concave if it is convex concave and for every $x \times y \in \mathbb{R}^n \times \mathbb{R}^m$ either $L(\cdot; y)$ is strictly convex or $L(x; \cdot)$ is strictly concave.

At the center of our definition of HCC games is a convex concave utility function L. Additionally, each player of the game is equipped with a set of operator functions. The minimization player is equipped with n functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ while the maximization player is equipped with m functions $g_j : \mathbb{R}^{m_j} \to \mathbb{R}$. We will assume in the rest of our discussion that f_i, g_j, L are all C^2 functions. The inputs $\theta_i \in \mathbb{R}^{n_i}$ and $\phi_j \in \mathbb{R}^{m_j}$ are grouped in two vectors

$$\theta = \begin{bmatrix} \theta_1 & \cdots & \theta_n \end{bmatrix}^\top \quad F(\theta) = \begin{bmatrix} f_1(\theta_1) & \cdots & f_n(\theta_n) \end{bmatrix}^\top$$
$$\phi = \begin{bmatrix} \phi_1 & \cdots & \phi_m \end{bmatrix}^\top \quad G(\phi) = \begin{bmatrix} g_1(\phi_1) & \cdots & g_m(\phi_m) \end{bmatrix}^\top$$

We are ready to define the hidden convex concave game

$$(\theta^*; \phi^*) = \arg\min_{\theta \in \mathbb{R}^N} \arg\max_{\phi \in \mathbb{R}^M} L(F(\theta); G(\phi)):$$

where $N = \sum_{i=1}^n n_i$ and $M = \sum_{j=1}^m m_j$. Given a convex concave function L, all stationary points of L are (global) Nash equilibria of the min-max game. We will call the set of all equilibria of L, von Neumann solutions of L and denote them by Solution(L). Unfortunately, Solution(L) can be empty for games defined over the entire $\mathbb{R}^n \times \mathbb{R}^m$. For games defined over convex compact sets, the existence of at least one solution is guaranteed by von Neumann's minimax theorem. Our definition of HCC games can capture games on restricted domains by choosing appropriately bounded functions f_i and g_j. In the following sections, we will just assume that Solution(L) is not empty. We note that our results hold for both bounded and unbounded f_i and g_j. We are now ready to write down the equations of the GDA dynamics for a HCC game:

$$\begin{aligned} \dot{\theta}_i &= -\nabla_{\theta_i} L(F(\theta); G(\phi)) &&= -\nabla_{\theta_i} f_i(\theta_i) \frac{\partial L}{\partial f_i}(F(\theta); G(\phi)) \\ \dot{\phi}_j &= \nabla_{\phi_j} L(F(\theta); G(\phi)) &&= \nabla_{\phi_j} g_j(\phi_j) \frac{\partial L}{\partial g_j}(F(\theta); G(\phi)) \end{aligned} \tag{1}$$

2.3 Reparametrization

The following lemma is useful in studying the dynamics of hidden games.

Lemma 1. Let $k : \mathbb{R}^d \to \mathbb{R}$ be a C^2 function. Let $h : \mathbb{R} \to \mathbb{R}$ be a C^1 function and $x(t)$ denote the unique solution of the dynamical system Σ_1. Then the unique solution for dynamical system Σ_2 is $z(t) = x(\int_0^t h(s)ds)$

$$\left\{ \begin{array}{rcl} \dot{x} & = & \nabla k(x) \\ x(0) & = & x_{init} \end{array} \right. : \Sigma_1 \qquad \left\{ \begin{array}{rcl} \dot{z} & = & h(t)\nabla k(z) \\ z(0) & = & x_{init} \end{array} \right. : \Sigma_2 \tag{2}$$

By choosing $h(t) = -\partial L(F(t); G(t))/\partial f_i$ and $h(t) = \partial L(F(t); G(t))/\partial g_j$ respectively, we can connect the dynamics of each θ_i and ϕ_j under Equation (1) to gradient ascent on f_i and g_j. Applying Lemma 1, we get that trajectories of θ_i and ϕ_j under Equation (1) are restricted to be subsets of the corresponding gradient ascent trajectories with the same initializations. For example, in Figure 1 $\theta_i(t)$ can not escape the purple section if it is initialized at (a) neither the orange section if it is initialiazed at (f). This limits the attainable values that $f_i(t)$ and $g_j(t)$ can take for a specific initialization. Let us thus define the following:

Definition 2. For each initialization $x(0)$ of Σ_1, $\mathrm{Im}_k(x(0))$ is the image of $k \circ x : \mathbb{R} \to \mathbb{R}$.

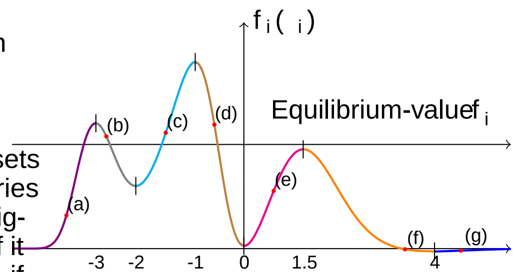

Figure 1: Neither Gradient Descent nor Ascent can traverse stationary points. An immediate consequence of Lemma 1 is that if we initialize in the above example $\theta_i(0)$ at (a), $f_i(\theta_i(t))$ can not escape the purple section. This extends to cases where θ_i is vector of variables.

Applying Definition 2 in the above example, $\mathrm{Im}_{f_i}(\theta_i(0)) = (f_i(-2), f_i(-1))$ if θ_i is initialized at (c). Additionally, observe that in each colored section $f_i(\theta_i(t))$ uniquely identifies $\theta_i(t)$. Generally, even in the case that θ_i are vectors, Lemma 1 implies that for a given $\theta_i(0)$, $f_i(\theta_i(t))$ uniquely identifies $\theta_i(t)$. As a result we get that a new dynamical system involving only f_i and g_j

Theorem 1. For each initialization $(\theta(0), \phi(0))$ of Equation (1), there are C^1 functions $X_{\theta_i(0)}, X_{\phi_j(0)}$ such that $\theta_i(t) = X_{\theta_i(0)}(f_i(t))$ and $\phi_j(t) = X_{\phi_j(0)}(g_j(t))$. If $(\theta(t), \phi(t))$ satisfy Equation (1) then $f_i(t) = f_i(\theta_i(t))$ and $g_j(t) = g_j(\phi_j(t))$ satisfy

$$\begin{aligned} \dot{f}_i &= -\|\nabla_{\theta_i} f_i(X_{\theta_i(0)}(f_i))\|^2 \frac{\partial L}{\partial f_i}(F, G) \\ \dot{g}_j &= \|\nabla_{\phi_j} g_j(X_{\phi_j(0)}(g_j))\|^2 \frac{\partial L}{\partial g_j}(F, G) \end{aligned} \tag{3}$$

By determining the ranges of f_i and g_j, an initialization clearly dictates if a von Neumann solution is attainable. In Figure 1 for example, any point of the pink, orange or blue colored section like (e), (f) or (g) can not converge to a von Neumann solution with $f_i(\theta_i) = f_i$. The notion of *safety* captures which initializations can converge to a given element of Solution(L).

Definition 3. We will call the initialization $(\theta(0), \phi(0))$ safe for a $(p, q) \in$ Solution(L) if $\theta_i(0)$ and $\phi_j(0)$ are not stationary points of f_i and g_j respectively and $p_i \in \mathrm{Im}_{f_i}(\theta_i(0))$ and $q_j \in \mathrm{Im}_{g_j}(\phi_j(0))$.

Leveraging the Center-Stable Manifold Theorem [55], the following observation shows that under mild assumptions almost all initializations are safe:

Theorem 2. If f_i and g_j have isolated stationary points, only strict saddle points, compact sublevel-sets, both equilibria $p_i \in (\max \mathrm{LocalMin}(f_i), \min \mathrm{LocalMax}(f_i))$ and $q_j \in (\max \mathrm{LocalMin}(g_j), \min \mathrm{LocalMax}(g_j))$, then almost all initializations are safe for $(p, q) \in$ Solution(L).

Finally, in the following sections we use some fundamental notions of stability. We call an equilibrium x^* of an autonomous dynamical system $\dot{x} = D(x(t))$ *stable* if for every neighborhood U of x^* there is a neighborhood V of x^* such that if $x(0) \in V$ then $x(t) \in U$ for all $t \geq 0$. We call a set S *asymptotically stable* if there exists a neighborhood R such that for any initialization $x(0) \in R$, $x(t)$ approaches S as $t \to +\infty$. If R is the whole space the set *globally asymptotically stable*.

3 Learning in Hidden Convex Concave Games

3.1 General Case

Our main results are based on designing a Lyapunov function for the dynamics of Equation (3):

Lemma 2. If L is convex concave and $(\theta(0), \phi(0))$ is a safe for $(p, q) \in$ Solution(L), then the following quantity is non-increasing under the dynamics of Equation (3):

$$H(F, G) = \sum_{i=1}^{N} \int_{p_i}^{f_i} \frac{z - p_i}{\|\nabla f_i(X_{\theta_i(0)}(z))\|^2} dz + \sum_{j=1}^{M} \int_{q_j}^{g_j} \frac{z - q_j}{\|\nabla g_j(X_{\phi_j(0)}(z))\|^2} dz \tag{4}$$

Observe that our Lyapunov function here is not the distance to (p, q) as in a classical convex concave game. The gradient terms account for the non constant multiplicative terms in Equation (3). Indeed if the game was not hidden and f_i and g_j were the identity functions then H would coincide with the Euclidean distance to (p, q). Our first theorem employs the above Lyapunov function to show that (p, q) is stable for Equation (3).

Theorem 3. If L is convex concave and $(\theta(0), \phi(0))$ is a safe for $(p, q) \in$ Solution(L), then (p, q) is stable for Equation (3).

Figure 2: Level sets of Lyapunov function of Equation (4) for both F and G being one dimensional sigmoid functions.

Clearly, for the special case of globally invertible functions F, G we could come up with an equivalent Lyapunov function in the θ,ϕ-space. In this case it is straightforward to transfer the stability results from the induced dynamical system of F, G (Equation (3)) to the initial dynamical system of θ,ϕ (Equation (1)). For example we can prove the following result:

Theorem 4. If f_i and g_j are sigmoid functions and L is convex concave and there is a $(\theta(0), \phi(0))$ that is safe for $(p, q) \in \text{Solution}(L)$, then $(F^{-1}(p), G^{-1}(q))$ is stable for Equation (1).

In the general case though, stability may not be guaranteed in the parameter space of Equation (1). We will instead prove a weaker notion of stability, which we call hidden stability. Hidden stability captures that if $(F(\theta(0)), G(\phi(0)))$ is close to a von Neumann solution, then $(F(\theta(t)), G(\phi(t)))$ will remain close to that solution. Even though hidden stability is weaker, it is essentially what we are interested in, as the output space determines the utility that each player gets. Here we provide sufficient conditions for hidden stability.

Theorem 5 (Hidden Stability)**.** Let $(p, q) \in \text{Solution}(L)$. Let R_{f_i} and R_{g_j} be the set of regular values[1] of f_i and g_j respectively. Assume that there is a $\delta > 0$ such that $[p_i - \delta, p_i + \delta] \subseteq R_{f_i}$ and $[q_j - \delta, q_j + \delta] \subseteq R_{g_j}$. Define

$$r(t) = \|F(\theta(t)) - p\|^2 + \|G(\phi(t)) - q\|^2.$$

If f_i and g_j are proper functions[2], then for every $\epsilon > 0$, there is an $\delta > 0$ such that

$$r(0) < \delta \implies \forall t \geq 0 : r(t) < \epsilon.$$

Unfortunately hidden stability still does not imply convergence to von Neumann solutions. [65] studied hidden bilinear games and proved that $\dot{H} = 0$ for this special class of HCC games. Hence, a trajectory is restricted to be a subset of a level set of H which is bounded away from the equilibrium as shown in Figure 2. To sidestep this, we will require in the next subsection the hidden game to be strictly convex concave.

3.2 Hidden strictly convex concave games

In this subsection we focus on the case where L is a strictly convex concave function. Based on Definition 1, a strictly convex concave game is not necessarily strictly convex strictly concave and thus it may have a continum of von Neumann solutions. Despite this, LaSalle's invariance principle, combined with the strict convexity concavity, allows us to prove that if $(\theta(0), \phi(0))$ is safe for $Z \subseteq \text{Solution}(L)$ then Z is locally asymptotically stable for Equation (3).

Lemma 3. Let L be strictly convex concave and $Z \subseteq \text{Solution}(L)$ is the non empty set of equilbria of L for which $(\theta(0), \phi(0))$ is safe. Then Z is locally asymptotically stable for Equation (3).

The above lemma however does not suffice to prove that for an arbitrary initialization $(\theta(0), \phi(0))$, $(F(t), G(t))$ approaches Z as $t \to +\infty$. In other words, a-priori it is unclear if $(F(\theta(0)), G(\phi(0)))$ is necessarily inside the region of attraction (ROA) of Z. To get a refined estimate of the ROA of Z, we analyze the behavior of H as f_i and g_j approach the boundaries of $\text{Im} f_i(\theta_i(0))$ and $\text{Im} g_j(\phi_j(0))$ and more precisely we show that the level sets of H are bounded. Once again the corresponding analysis is trivial for convex concave games, since the level sets are spheres around the equilibria.

Theorem 6. Let L be strictly convex concave and $Z \subseteq \text{Solution}(L)$ is the non empty set of equilbria of L for which $(\theta(0), \phi(0))$ is safe. Under the dynamics of Equation (1) $(F(\theta(t)), G(\phi(t)))$ converges to a point in Z as $t \to \infty$.

The theorem above guarantees convergence to a von Neumann solution for all initializations that are safe for at least one element of Solution(L). However, this is not the same as global asymptotic stability. To get even stronger guarantees, we can assume that all initializations are safe. In this case it is straightforward to get a global asymptotic stability result:

Corollary 1. Let L be strictly convex concave and assume that all intitializations are safe for at least one element of Solution(L). The following set is globally asymptotically stable for continuous GDA dynamics.

$$\{(\theta, \phi) \in \mathbb{R}^n \times \mathbb{R}^m : (F(\theta), G(\phi)) \in \text{Solution}(L)\}$$

[1] A value $a \in \text{Im} f$ is called a regular value of f if $\forall q \in \text{dom} f : f(q) = a$, it holds $\nabla f(q) \neq 0$.

[2] A function is proper if inverse images of compact subsets are compact.

Notice that the above approach on global asymptotic convergence using Lyapunov arguments can be extended to other popular alternative gradient-based heuristics like variations of Hamiltonian Gradient Descent. For concision, we defer the exact statements, proofs in the supplement.

3.3 Convergence via regularization

Regularization is a key technique that works both in the practice of GANs [17, 33] and in the theory of convex concave games [56, 59, 60]. Our settings of hidden convex concave games allows for provable guarantees for regularization in a wide class of settings, bringing closer practical and theoretical guarantees. Let us have a utility $L(x; y)$ that is convex concave but not strictly. Here we will propose a modified utility L^0 that is strictly convex strictly concave. Specifically we will choose

$$L^0(x; y) = L(x; y) + \frac{}{2}\|x\|^2 \quad \frac{}{2}\|y\|^2$$

The choice of the parameter captures the trade-off between convergence to the original equilibrium of L and convergence speed. On the one hand, invoking the implicit function theorem, we get that for small the equilibria of L are not significantly perturbed.

Theorem 7. If L is a convex concave function with invertible Hessians at all its equilibria, then for each > 0 there is a > 0 such that L^0 has equilibria that are -close to the ones of L.

Note that invertibility of the Hessian means that L must have a unique equilibrium. On the other hand increasing increases the rate of convergence of safe initializations to the perturbed equilibrium.

Theorem 8. Let $((0); (0))$ be a safe initialization for the unique equilibrium of L^0 $(p; q)$. If

$$r(t) = \|F((t)) \quad p\|^2 + \|G((t)) \quad q\|^2$$

then there are initialization dependent constants $c_0; c_1 > 0$ such that $r(t) \quad c_0 \exp(\quad c_1 t)$.

4 Applications

In this section, we discuss how HCC framework can be used to give new insights in a variety of application areas including min-max training for GANs and Evolutionary Game Theory. We also describe applications of regularization to normal form zero sum games in Appendix D.3.

Figure 3: Both the ℓ_2 distance from the equilibrium and $H(t)$ converge to zero but only the latter does so monotonically. For p_{data} we choose a fully mixed distribution of dimension $= 4$. Given the sigmoid activations all the initializations are safe. We defer the detailed proof of convergence in Appendix D.2.

Hidden strictly convex-concave games. We will start our discussion with the fundamental generative architecture of [26]'s GAN. In the vanilla GAN architecture, as it is commonly referred, our goal is to find a generator distribution p_G that is close to an input data distribution p_{data}. To find such a generator function, we can use a discriminator D that "criticizes" the deviations of the generator from the input data distribution. For the case of a discrete p_{data} over a set N, the minimax problem of [26] is the following:

$$\min_{\substack{p_G(x) \quad 0; \\ \sum_{x \in N} p_G(x)=1}} \quad \max_{D \in (0;1)^{|N|}} V(G; D)$$

where $V(G; D) = \sum_{x \in N} p_{data}(x) \log(D(x)) + \sum_{x \in N} p_G(x) \log(1 \quad D(x))$. The problem above can be formulated as a constrained strictly convex-concave hidden game. On the one hand, for a fixed discriminator D, the $V(G; D)$ is linear over the $p_G(x)$. On the other hand, for a fixed generator G,

$V(G_\theta; D_\phi)$ is strongly-concave. We can implement the inequality constraints on both the generator probabilities and discriminator using sigmoid activations. For the equality constraint $\sum_{x \in N} p_G(x) = 1$ we can introduce a Langrange multiplier. Having effectively removed the constraints, we can see in Figure 3, the dynamics of Equation (1) converge to the unique equilibrium of the game, an outcome consistent with our results in Corollary 1. While the Euclidean distance to the equilibrium is not monotonically decreasing, $H(t)$ is.

Hidden convex-concave games & Regularizaiton. An even more interesting case is Wassertein GANs–WGANs [4]. One of the contributions of [36] is to show that WGANs trained with Stochastic GDA can learn the parameters of Gaussian distributions whose samples are transformed by non-linear activation functions. It is worth mentioning that the original WGAN formulation has a Lipschitz constraint in the discriminator function. For simplicity, [36] replaced this constraint with a quadratic regularizer. The min-max problem for the case of one-dimensional Gaussian $N(0, \sigma^2)$ and linear discriminator $D_v(x) = v^\top x$ with x^2 activation is:

$$
\begin{aligned}
\min_{\theta \in R} \max_{v \in R} V_{WGAN}(G_\theta; D_v) &= E_{X \sim p_{data}}[D(X)] - E_{X \sim p_G}[D(X)] - v^2/2 \\
&= E_{x \sim N(0;\sigma^2)^2}[vx] - E_{x \sim N(0;\theta^2)^2}[vx] - v^2/2 \\
&= (\sigma^2 - \theta^2)v - v^2/2
\end{aligned}
$$

Observe that V_{WGAN} is not convex-concave but it can posed as a hidden strictly convex-concave game with $G(\theta) = (\sigma^2 - \theta^2)$ and $F(v) = v$. When computing expectations analytically without sampling, Theorem 6 guarantees convergence. In contrast, without the regularizer, V_{WGAN} can be modeled as a hidden bilinear game and thus GDA dynamics cycle. Empirically, these results are robust to discrete and stochastic updates using sampling as shown in Figure 4. Therefore regularization in the work of [36] was a vital ingredient in their proof strategy and not just an implementation detail.

Figure 4: On the left, we show the trajectories of regularized GDA for $\sigma^2 = 1$ as well as the level sets of Equation (4). All trajectories (green curves-initialized at the red points) converge to one of the two equilibria $(0; 1)$ and $(0; -1)$ whereas without regularization, GDA would cycle on the level sets. In the right figure, we replace the exact expectations in V_{WGAN} with approximations via sampling and continuous time updates on θ and v with discrete ones. For small learning rates and large sample sizes, unregularized GDA continues to cycle. In contrast, the regularization approach of [36] converges to the $(0; 1)$ equilibrium.

The two applications of HCC games in GANs are not isolated findings but instances of a broader pattern that connects HCC games and standard GAN formulations. As noted by [27], if updates in GAN applications were directly performed in the "functional space", i.e. the generator and discriminator outputs, then standard arguments from convex concave optimization would imply convergence to global Nash equilibria. Indeed, standard GAN formulations like the vanilla GAN [26], f-GAN [50] and WGAN [4] can all be thought of as convex concave games in the space of generator and discriminator outputs. Given that the connections between convex concave games and standard GAN objectives in the output space is missing from recent literature, in Appendix D.1 we show how one can apply Von Neumann's minimax theorem to derive the optimal generators and discriminators even in the non-realizable case. In practice, the updates happen in the parameter space and thus convexity arguments no longer apply. Our study of HCC games is a stepping stone towards bridging the gap in convergence guarantees between the case of direct updates in the output space and the parameter space.

Figure 5: The above figures describes the evolution of the expected phenotype for two species A,B. The left one corresponds to a safe initialization leading to periodic trajectories. The middle one corresponds to an unsafe initialization where $_1(0) = {}_2(0)$. The dynamics converge albeit to a spurious equilibrium, that is different from the hidden game equilibrium (dash lines). Finally, the right one corresponds to a safe initialization of the regularized game, which converges to a slightly perturbed equilibrium (Theorem 7).

Evolutionary Game Theory & Biology. The study of learning dynamics in games has always been strongly and inherently connected with mathematical models of biology and evolution. Typically, this line of research is studied under the name of *Evolutionary Game Theory* [28, 67]. Zero-sum games and variants thereof are of particular interest for this line of work as they encode settings of direct competition between species (e.g., prey-predator or host-parasite/virus). Even in the simplest such setting of matrix zero-sum games, used to capture competition between asexually reproducing species, it is well known that the emerging dynamics can be non-equilibrating and even chaotic [61, 58].

Studying the effects of evolutionary competition between sexually evolving species results in significantly more intricate models, as it does not suffice to merely keep track of the fractions of the different types of individuals that self-replicate. Instead it is necessary to keep a much more detailed accounting of the evolution of the frequencies of different genes that get reshuffled and recombined to create new individuals, whilst giving evolutionary preference to the most fit individuals given the current environment. Recent work on intersection of learning theory and game theory has provided concrete such game theoretic models [37, 14, 44, 42]. Due to the intricate nature of their dynamics, deciding even the simplest questions in regards to them (e.g. does genetic diversity survive or not?) is typically computationally hard [43].

A notable exception, where the dynamics of sexual evolution and, in fact, sexual competition have been relatively thoroughly understood, is the work of [57] on two species (host-parasite) antagonism. The outcome of this competition depends on their respective phenotypes (informally their properties, e.g., large wings versus small wings.) of the two species. The crucial assumption that makes this model theoretically tractable is that the phenotype for each species is a Boolean attribute (this assumption is also used [38]). Despite these simplifications, the dynamics are still not equilibrating and are, in fact, cyclic for almost all initial conditions. Two natural questions emerge: 1) Is the almost everywhere condition necessary? I.e. Do there exist initial conditions which are not cyclic? 2) More importantly, can a slightly perturbed dynamic stabilize these systems and converge to a meaningful equilibrium? Next, we will see how our framework addresses both of these questions.

To understand the connection these we will examine the model of [57] in more detail. Concretely, the phenotype of species A, B can be described as a Boolean function over the species genome which is encoded by a binary string (this acts as a simplified version of a DNA string). While the *phenotype* plays the dominant role for the survival of the species, sexual reproduction modifies only the *genotype* of an organism. As a result the species are actually involved in a hidden zero-sum game. More formally, each species is game-theoritically represented as a team of agents where each agent controls one bit of the genotype:

$$G_A = (g_1^A, \ldots, g_n^A), \; G_B = (g_1^B, \ldots, g_m^B)$$
$$u_A = L[\text{Phenotype}_A(G_A), \text{Phenotype}_B(G_B)]$$
$$u_B = -u_A$$

Where $g_i^A, g_j^B \in \{0, 1\}$, $\text{Phenotype}_A, \text{Phenotype}_B$ is a Boolean function (e.g., AND, XOR) and L is a 2×2 matrix encoding a zero-sum game (e.g., Matching Pennies). Naturally, one can allow agents to use randomized/mixed strategies in which case the expected utilities of all agents/genes are defined using the standard multi-linear extension of utilities. Thus, these models of evolutionary sexual competition share the same basic structure as hidden linear-linear games, which explains their recurrent, non-equilibrating nature [65].

In Figure 5, each gene/agent g_i^A tunes one real variable $_i$ such that $\Pr[g_i^A = 1] = \ (\ _i)$ and gene/agent g_j^B tunes one real variable $_j$ correspondingly. Choosing as Boolean phenotype to be the XOR of two genes, almost all initializations are safe for any bilinear game with a mixed equilibrium. Actually, only the case $_1(0) = \ _2(0)$ or $_1(0) = \ _2(0)$ can be problematic, since for XOR the expected phenotype is bounded in $[0, 0.5]$ and a mixed equilibrium out of this range would be infeasible. Finally, leveraging Theorem 7, we can design a regularized version of the game such that the dynamics converge arbitrarily close to the true von Neumann solution of these games, which is encoded by the min-max strategies of the hidden bi-linear zero-sum game.

5 Discussion & Future Work

While this work is a promising rst step towards understanding GAN training, signi cant challenges remain. Neural network architectures do not use disjoint set of parameters for each of the outputs. Additionally, the hidden competition of GANs can take place in an output space of probability distributions and classi ers whose vector space dimension is typically in nite. On the bright side, we establish point-wise (day to day) convergence results which are, to the best of our knowledge, the rst result of their kind for a wide class of non-convex non-concave games that do not necessarily satisfy the Polyak- ojasiewicz conditions studied in [68]. Such conditions imply that the notions of saddle points, global min-max and stationary points coincide. Instead our work showcases how to make progress without leveraging such strong assumptions in zero-sum games. Beyond ML applications, we believe that our framework could provide even further insights for evolutionary game theory, mathematical biology as well as team-zero-sum games. For example an interesting hybrid class of games could be network generalizations of team-zero-sums games, e.g. by combining [12] and [57].

Acknowledgments and Disclosure of Funding

This research/project is supported in part by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016), NRF 2018 Fellowship NRF-NRFF2018-07, NRF2019-NRF-ANR095 ALIAS grant, grant PIE-SGP-AI-2018-01, AME Programmatic Fund (Grant No. A20H6b0151) from the Agency for Science, Technology and Research (A*STAR). Additionally, E.V. Vlatakis-Gkaragkounis is grateful to be supported by NSF grants CCF-1703925, CCF1763970, CCF-1814873, CCF-1563155, and by the Simons Collaboration on Algorithms and Geometry. He would like to acknowledge the support of Onassis Foundation under the Scholarship(ID: FZN 010-1/2017-2018.)

References

[1] Jacob Abernethy, Kevin A Lai, K r Y Levy, and Jun-Kun Wang. Faster rates for convex-concave games. In *COLT*, 2018.

[2] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv:1906.02027*, 2019.

[3] Leonard Adolphs, Hadi Daneshmand, Aurélien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *AISTATS*, 2019.

[4] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.

[5] James Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: vanishing regret with non-vanishing step sizes. In *NeurIPS*, 2019.

[6] James P Bailey, Gauthier Gidel, and Georgios Piliouras. Finite regret and cycles with xed step-size via alternating gradient descent-ascent. In *COLT*, 2020.

[7] James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *EC*, 2018.

[8] David Balduzzi, Sébastien Racanière, James Martens, Jakob N. Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *ICML*, 2018.

[9] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999.

[10] Sanjay P. Bhat and Dennis S. Bernstein. Nontangency-based lyapunov tests for convergence and stability in systems having a continuum of equilibria. *SIAM J. Control Optim*, 2003.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004.

[12] Yang Cai and Costantinos Daskalakis. On minmax theorems for multiplayer games. In *SODA*, SODA, 2011.

[13] Nikolo Cesa-Bianchi and Gabor Lugoisi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[14] Erick Chastain, Adi Livnat, Christos Papadimitriou, and Umesh Vazirani. Algorithms, games, and evolution. *PNAS*, 2014.

[15] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *COLT*, 2019.

[16] Yun Kuen Cheung and Georgios Piliouras. Chaos, extremism and optimism: Volume analysis of learning in games. In *NeurIPS*, 2020.

[17] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR*, 2018.

[18] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *NeurIPS*, 2018.

[19] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, 2019.

[20] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. *arxiv:2009.09623*, 2020.

[21] Lawrence C Evans. Partial differential equations and monge-kantorovich mass transfer. *Current developments in mathematics*, 1997.

[22] Ky Fan. Minimax theorems. *PNAS*, 1953.

[23] Farzan Farnia and Asuman E. Ozdaglar. Do gans always have nash equilibria? In *ICML*, 2020.

[24] Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *ICML*, 2020.

[25] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[27] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arxiv:1701.00160*, 2017.

[28] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998.

[29] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, 2019.

[30] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS*, 2020.

[31] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arxiv:1902.00618*, 2019.

[32] Hassan K Khalil. *Nonlinear systems; 3rd ed.* Prentice-Hall, 2002.

[33] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans. In *ICML*, 2019.

[34] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arxiv:1710.07406*, 2017.

[35] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *COLT*, 2016.

[36] Qi Lei, Jason Lee, Alex Dimakis, and Constantinos Daskalakis. SGD learns one-layer networks in wgans. In *ICML*, 2020.

[37] Adi Livnat, Christos Papadimitriou, Jonathan Dushoff, and Marcus W. Feldman. A mixability theory for the role of sex in evolution. *PNAS*, 2008.

[38] Adi Livnat, Christos Papadimitriou, Aviad Rubinstein, Andrew Wan, and Gregory Valiant. Satisfiability and evolution. In *FOCS*, 2014.

[39] Chris J. Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arxiv:1809.05042*, 2018.

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[41] Eric Mazumdar and Lillian J Ratliff. Local nash equilibria are isolated, strict local nash equilibria in 'almost all' zero-sum continuous games. In *Proc. IEEE Conf. Decis. Control*, 2019.

[42] Ruta Mehta, Ioannis Panageas, and Georgios Piliouras. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. In *ITCS*, 2015.

[43] Ruta Mehta, Ioannis Panageas, Georgios Piliouras, and Sadra Yazdanbod. The Computational Complexity of Genetic Diversity. *European Symposium on Algorithms (ESA)*, 2016.

[44] Reshef Meir and David Parke. A note on sex, evolution, and the multiplicative updates algorithm. In *AAMAS*, 2015.

[45] 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. In *ICLR*, 2019.

[46] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA*, 2018.

[47] Lars M. Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018.

[48] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, 2020.

[49] Yurii E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

[50] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016.

[51] Brendan O'Donoghue and Chris J. Maddison. Hamiltonian descent for composite objectives. In *NeurIPS*, 2019.

[52] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *ITCS*, 2017.

[53] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *ITCS*, 2017.

[54] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *NeurIPS*, 2019.

[55] Lawrence Perko. *Differential Equations and Dynamical Systems*. Springer, 3nd. edition, 1991.

[56] Julien Pérolat, Rémi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro A. Ortega, Neil Burch, Thomas W. Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. *arxiv:2002.08456*, 2020.

[57] Georgios Piliouras and Leonard J Schulman. Learning dynamics and the co-evolution of competing sexual species. In *ITCS*, 2018.

[58] Georgios Piliouras and Jeff S. Shamma. Optimization despite chaos: Convex relaxations to complex limit sets via poincaré recurrence. In *SODA*, 2014.

[59] Kevin Roth, Aurélien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NeurIPS*, 2017.

[60] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D. Lee. On the convergence and robustness of training gans with regularized optimal transport. In *NeurIPS*, 2018.

[61] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. Chaos in learning a simple two-person game. *PNAS*, 2002.

[62] Leonard Schulman and Umesh V Vazirani. The duality gap for two-team zero-sum games. In *ITCS*, 2017.

[63] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 2017.

[64] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[65] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS*, 2019.

[66] Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. In *NeurIPS*, 2018.

[67] Jörgen Weibull. *Evolutionary Game Theory*. MIT Press; Cambridge, MA: Cambridge University Press., 1995.

[68] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *NeurIPS*, 2020.

[69] Guojun Zhang, Pascal Poupart, and Yaoliang Yu. Optimality and Stability in Non-Convex-Non-Concave Min-Max Optimization. *arXiv:2002.11875*, 2020.

[70] Guojun Zhang and Yaoliang Yu. Convergence of gradient methods on bilinear zero-sum games. In *ICLR*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not see any ethical or future societal consequences of this work, since its motivation is strongly theoretical.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results?[Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?[N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The presented experiments are used for illustrative purposes and only to validate the theoretical findings which are the core results of this work.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Background

A.1 Background in dynamical systems

Our analysis combines tools from dynamical systems, stability analysis and invariance principles theory. We start with the definitions of the different stability notions. We remind the well known Lyapunov's Lyapunov stability criterion (Theorem 9) Stability analysis in convex concave games is further complicated due to the possibility of non-isolated fixed points. To tackle this issue, we recall Krasovskii-LaSalle's Invariance Principle (Theorem 10), a powerful result that has several implications for the asymptotic stability of a set in an autonomous (possibly nonlinear) dynamical system. In the special case where the goal set contains only stable fixed points a pointwise convergence theorem can be derived (Theorem 11). The Center-Stable Manifold Theorem (Theorem 12) is going to be a key ingredient of the proof of Theorem 2. Finally, we remind the notions of diffeomorphism and topological conjugacy of two dynamical systems, which are useful to transfer behavioral claims between equivalent dynamics.

Let $f : D \to \mathbb{R}^n$ be a locally Lipschitz map from a domain $D \subseteq \mathbb{R}^n$ to $\mathbb{R}^n$. We consider dynamical systems of the form

$$\dot{x} = f(x) \qquad (?)$$

A point x for which $f(x) = 0$ is called a fixed point. We will be interested in the following notions of stability for the fixed point points of Equation (?).

Definition 4 (Stability properties, [32, Definition 4.1]) The fixed point $x = 0$ of Equation (?) is

- stable if, for each $\epsilon > 0$, there is a $\delta = \delta(\epsilon) > 0$ such that

$$\|x(0)\| < \delta \implies \|x(t)\| < \epsilon \quad \forall t \geq 0$$

- unstable if it is not stable
- asymptotically stable if it is stable and δ can be chosen such that

$$\|x(0)\| < \delta \implies \lim_{t \to \infty} x(t) = 0$$

The Lyapunov Theorem will be a useful tool to prove (asymptotic) stability of a fixed point.

Theorem 9 (Lyapunov Theorem, [32, Theorem 4.1]) Let $x = 0$ be a fixed point for Equation (?) and $D \subseteq \mathbb{R}^n$ be a domain containing $x = 0$. Let $V : D \to \mathbb{R}$ be a continuously differentiable function such that

$$V(0) = 0 \text{ and } V(x) > 0 \text{ in } D - \{0\}$$
$$\dot{V}(x) \leq 0 \text{ in } D$$

then $x = 0$ is stable. Moreover if

$$\dot{V}(x) < 0 \text{ in } D - \{0\}$$

then $x = 0$ is asymptotically stable.

Unfortunately, the Lyapunov theorem is not very helpful when it comes to proving convergence in dynamical systems with non isolated fixed points. By definition, non-isolated fixed points cannot be asymptotically stable. Non isolated fixed points may give rise to more complex behavior than point-wise convergence.

Definition 5. We say that a trajectory $x(t)$ approaches a set M as $t \to \infty$ if for each $\epsilon > 0$ there is a $T > 0$ such that

$$\operatorname{dist}(x(t), M) < \epsilon, \quad \forall t > T$$

where the operator "dist" is the minimum distance from a point to a set M

$$\operatorname{dist}(p, M) = \inf_{x \in M} \|p - x\|$$

Definition 6. We say that a set M is invariant for Equation (?) if

$$x(0) \in M \implies x(t) \in M, \quad \forall t \in \mathbb{R}$$

We will say M is positively invariant if the above holds for $t \geq 0$.

We are ready to state LaSalle's Invariance Principle, a general theorem that can help us study the stability of non isolated fixed points.

Theorem 10 (LaSalle's Invariance Principle, [32, Theorem 4.4]). Let $\Omega \subset D$ be a compact set that is positively invariant with respect to Equation (?). Let $V : D \to \mathbb{R}$ be a continuously differentiable function such that $\dot{V}(x) \leq 0$ in Ω. Let E be the set of all points where $\dot{V}(x) = 0$. Let M be the largest invariant set in E. Then every solution starting in Ω approaches M as $t \to \infty$.

LaSalle's theorem does not give us pointwise convergence directly. But in the special case that M contains only stable fixed points we can apply the following theorem

Theorem 11 (Pointwise Convergence Theorem, [10, Proposition 5.4]). Let $x(t)$ be a trajectory of Equation (?). If the positive limit sets of $x(t)$ contain a stable fixed point then $x(t)$ converges to it as $t \to \infty$.

Definition 7 (Differomorphism, [55]). Let U, V be manifolds. A map $f : U \to V$ is called a diffeomorphism if f carries U onto V and also both f and f^{-1} are smooth.

Definition 8 (Topological conjugacy, [55]). Two flows $\phi_t : A \to A$ and $\psi_t : B \to B$ are conjugate if there exists a homeomorphism $g : A \to B$ such that

$$\forall \mathbf{x} \in A, t \in \mathbb{R} : g(\phi_t(\mathbf{x})) = \psi_t(g(\mathbf{x}))$$

Furthermore, two flows $\phi_t : A \to A$ and $\psi_t : B \to B$ are diffeomorphic if there exists a diffeomorphism $g : A \to B$ such that

$$\forall \mathbf{x} \in A, t \in \mathbb{R} : g(\phi_t(\mathbf{x})) = \psi_t(g(\mathbf{x})).$$

If two flows are diffeomorphic, then their vector fields are related by the derivative of the conjugacy. That is, we get precisely the same result that we would have obtained if we simply transformed the coordinates in their differential equations.

Theorem 12 (Stable Manifold Theorem for Continuous Time Dynamical Systems [55] p.120). Let E be an open subset of $\mathbb{R}^n$ containing the origin, let $f \in C^1(E)$, and let ϕ_t be the flow of the nonlinear system $\dot{\mathbf{x}} = f(\mathbf{x})$. Suppose that $f(0) = 0$ and that $Df(0)$ has k eigenvalues with negative real part and $n - k$ eigenvalues with positive real part. Then there exists a k-dimensional differentiable manifold S tangent to the stable subspace E^s of the linear system $\dot{\mathbf{x}} = Df(0)\mathbf{x}$ at 0 such that for all $t \geq 0$, $\phi_t(S) \subset S$ and for all $\mathbf{x}_0 \in S$:

$$\lim_{t \to \infty} \phi_t(\mathbf{x}_0) = 0$$

and there exists an $n - k$ dimensional differentiable manifold U tangent to the unstable subspace E^u of the linear system $\dot{\mathbf{x}} = Df(0)\mathbf{x}$ at 0 such that for all $t \leq 0$, $\phi_t(U) \subset U$ and for all $\mathbf{x}_0 \in U$:

$$\lim_{t \to -\infty} \phi_t(\mathbf{x}_0) = 0$$

Remark 1. While the focus of our work relies on results in continuous dynamics, for the interested reader the broader possible implications of our results in discrete time algorithms should be highlighted. Indeed, under some technical conditions described in work of Benaim [9], whose verification for HCC games lies beyond the scope of the current work, one could argue that discretized dynamics, even the stochastic ones, have the same convergence behaviour as their continuous counterparts. Nevertheless, we believe that our continuous-time results will serve as a fundamental building block of any argument in that direction. Indeed, our diverse set of simulations, which are discrete-time implementations, are in perfect agreement with our theoretical predictions showcasing the applicability of our results.

A.2 Background in convex optimization

> For the sake of completeness, we recall here the definition of (strict) convex/concave function and its first order necessary and sufficient criterion. We will also discuss strong convexity and its second order characterizations.

We will be interested in notions from convex optimization throughout this work

Definition 9 ([11, p. 67]). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function then

- f is convex if

$$\forall x, y \in \mathbb{R}^n, t \in [0, 1] : f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$$

- f is strictly convex if

$$\forall x, y \in \mathbb{R}^n, t \in (0, 1) : f(tx + (1 - t)y) < tf(x) + (1 - t)f(y)$$

- f is (strictly) concave if $-f$ is (strictly) convex.

We will also use the first order characterizations of convex and concave functions

Theorem 13 ([11, p. 69-70]). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function.

- f is convex if and only if $\forall x, y \in \mathbb{R}^n : f(y) \geq f(x) + \nabla f(x)^T(y - x)$
- f is concave if and only if $\forall x, y \in \mathbb{R}^n : f(y) \leq f(x) + \nabla f(x)^T(y - x)$

To establish convergence rates, we will use the notion of strong convexity

Definition 10 ([49, p. 63]). A continuously differentiable function f of $\mathbb{R}^n$ will be called μ-strongly convex for a positive constant μ if for all $x, y \in \mathbb{R}^n$ we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2$$

We will also use second order characterizations of strong convexity

Theorem 14 ([49, p. 65]). A twice continuously differentiable function f is μ-strongly convex for a positive constant μ if and only if for all $x \in \mathbb{R}^n$ we have

$$\nabla^2 f(x) \succeq \mu I$$

Symmetrically, a function will be called strongly concave if $-f$ is μ-strongly convex.

A.3 Background in Game Theory

> In this short section, we remind to the reader a generalization of Von-Neumann's Minimax theorem, which we will exploit to analyze the equilibrium solution of the different GANs' architectures. A special case of Fan's minimax theorem is the following

Corollary 2 (Fan's minimax theorem, [22]). Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be convex non-empty sets. Suppose that X is compact and $f : X \times Y \to \mathbb{R}$ is a function such that $f(\cdot; y)$ is lower semicontinuous on X for each $y \in Y$ and that f is convex concave. Then we have that

$$\min_{x \in X} \sup_{y \in Y} f(x; y) = \sup_{y \in Y} \min_{x \in X} f(x; y).$$

B Preliminaries

> The below time-reparametrization lemma shows that the solution for a non-autonomous system, multiplicative to a gradient flow can be derived by just time-rescaling of the solution of the simplified gradient ascent dynamics. Indeed, since the multiplicative term is common across all terms of the vector field then over the time it dictates only the magnitude of the vector field (the speed of the motion), but does not affect the directionality other than moving backwards or forwards along the same trajectory.

Lemma 1. Let $k : \mathbb{R}^d \to \mathbb{R}$ be a C^2 function. Let $h : \mathbb{R} \to \mathbb{R}$ be a C^1 function and $x(t)$ denote the unique solution of the dynamical system Σ_1. Then the unique solution for dynamical system Σ_2 is $z(t) = x(\int_0^t h(s)ds)$

$$\begin{cases} \dot{x} = \nabla k(x) \\ x(0) = x_{init} \end{cases} : \Sigma_1 \qquad \begin{cases} \dot{z} = h(t)\nabla k(z) \\ z(0) = x_{init} \end{cases} : \Sigma_2 \tag{5}$$

Proof. Firstly, notice that it holds $x(0) = x_{init}$ and $\dot{x} = \nabla k(x)$, since x is the unique solution of Σ_1. It is easy to check that:

$$z(0) = x\left(\int_0^0 h(s)ds\right) = x(0) = x_{init}$$

$$\dot{z} = \dot{x}\left(\int_0^t h(s)ds\right) \cdot \frac{d[\int_0^t h(s)ds]}{dt}$$

$$= \nabla k\left(x\left(\int_0^t h(s)ds\right)\right) \cdot h(t) = \nabla k(z)h(t)$$

□

Remark 2. Notice that Lemma 1 applies a-fortiori in the class of HCCs games. Indeed, it is important to observe that allowing $h(t)$ to be any smooth function of time already covers already the cases where $h(t)$ depends also on $z(t)$. The key intuition is that for the needs of our existential result, we can always assume that the exact solution of Σ_2 is known. Thus we can always substitute the $z(t)$-dependent $h(t)$ with an expression that depends only directly on t

Example 1. Let's consider the simplest example where $n = 1$ and $k(x) = x^2/2$ with for a state-dependent multiplicative factor $h(t) = z(t)$. Then,

- $\Sigma_1 \{ \dot{x}(t) = x(t); x(0) = 1 \}$, whose the unique solution is $x(t) = e^t$.
- $\Sigma_2 \{ \dot{z}(t) = h(t)z(t) = z(t)^2; z(0) = 1 \}$ whose unique solution is $z(t) = 1/(1-t)$.

We can still apply the aforementioned reparametrization lemma by substituting $h(t) = z(t)$ with $h(t) = 1/(1-t)$. In other words, any $h(x; t)$ can always be written given any initial condition as $h_x(t)$.

Remark 3. It is worth mentioning that due to the multiplicative factor $h(t)$ in Σ_2 and as a result the time reparamatrization factor $\int_0^t h(s)ds$ can take both positive or negative values based on the GDA dynamics. By Lemma 1, this allows the GDA dynamics to visit states that are visited by the simple GD dynamics for $t < 0$.

> In order to leverage the convex-concave properties of the operators in our hidden structure under the Gradient Descent Ascent dynamics we need to recover the equivalent system (T) in the operator space $\begin{bmatrix}\dot F\\ \dot G\end{bmatrix} = T\begin{bmatrix}F\\ G\end{bmatrix}$.
>
> $$(\cdot) := \begin{cases}\dot\theta = -\nabla_\theta L(F(\theta);G(\phi)) & \dot\theta_i = -\nabla_{\theta_i} f_i(\theta_i)\,\partial L(F(t);G(t))/\partial f_i\\ \dot\phi = \nabla_\phi L(F(\theta);G(\phi)) & \dot\phi_j = \nabla_{\phi_j} g_j(\phi_j)\,\partial L(F(t);G(t))/\partial g_j\end{cases}$$
>
> From this point, applying the aforementioned lemma, under GDA each f_i and g_j follows a time dependent rescaling of the corresponding gradient ascent solution. Exploiting the monotonicity of $f_i(t)$ and $g_j(t)$ under gradient ascent, we can construct an invertible map between the parameter space $\{\theta_i;\phi_j\}$ and the operator space $\{f_i;g_j\}$ which allows us to construct the equivalent system T in the operator space. Notice that the properties of gradient ascent are crucial since the operator space can be arbitrarily smaller in dimension. In this case a smooth invertible map that is common for all initializations cannot exist.

Theorem 1. For each initialization $(\theta(0);\phi(0))$ of Equation (1), there are C^1 functions $X_{\theta_i(0)}; X_{\phi_j(0)}$ such that $\theta_i(t) = X_{\theta_i(0)}(f_i(t))$ and $\phi_j(t) = X_{\phi_j(0)}(g_j(t))$. If $(\theta(t);\phi(t))$ satisfy Equation (1) then $f_i(t) = f_i(\theta_i(t))$ and $g_j(t) = g_j(\phi_j(t))$ satisfy

$$\begin{aligned}\dot f_i &= -\|\nabla_{\theta_i} f_i(X_{\theta_i(0)}(f_i))\|^2\frac{\partial L}{\partial f_i}(F;G)\\ \dot g_j &= \|\nabla_{\phi_j} g_j(X_{\phi_j(0)}(g_j))\|^2\frac{\partial L}{\partial g_j}(F;G)\end{aligned} \tag{3}$$

Proof. Let us first study a simpler dynamical system $(\cdot)$ with unique solution of $\cdot_{\theta_i(0)}(t)$.

$$(\cdot)\quad\begin{cases}\dot z = \nabla f_i(z)\\ z(0) = \theta_i(0)\end{cases}$$

It is easy to observe that:

$$\dot f = \nabla f(z)\dot z = \|\nabla f(z)\|^2$$

If $\theta_i(0)$ is a stationary point of f_i then the trajectory of z is a single point. But the trajectory of θ_i under the dynamics of Equation (1) is also a single point so we can pick the following function

$$X_{\theta_i(0)}(f_i) = \theta_i(0):$$

On the other hand if $\theta_i(0)$ is not a stationary point of f_i, f_i continuously increases along the trajectory of $(\cdot)$. Therefore $A_{\theta_i(0)}(t) = f_i(\cdot_{\theta_i(0)}(t))$ is an increasing function and therefore invertible. Let us call $A^{-1}_{\theta_i(0)}(f_i)$ the inverse.

Let's recall now the θ_i part of the dynamical system of interest Equation (1)

$$\dot\theta_i = -\nabla_{\theta_i} f_i(\theta_i)\frac{\partial L}{\partial f_i}(F(\theta);G(\phi))$$

initialized at $\theta_i(0)$. Applying Lemma 1 for the first equation with

$$h(t) = -\frac{\partial L}{\partial f_i}(F(\theta(t));G(\phi(t)))$$

we have that under the dynamics of Equation (1)

$$\theta_i(t) = \cdot_{\theta_i(0)}\left(\int_0^t h(s)ds\right) \tag{P}$$

Thus it holds

$$f_i(\theta_i(t)) = f\left(\cdot_{\theta_i(0)}\left(\int_0^t h(s)ds\right)\right) = A_{\theta_i(0)}\left(\int_0^t h(s)ds\right)$$

or equivalently

$$\int_0^t h(s)ds = A^{-1}_{\theta_i(0)}(f_i(\theta_i(t)))$$

Plugging in back to Equation (P)

$$\theta_i(t) = \psi_{\theta_i(0)}(A^{-1}_{\theta_i(0)}(f_i(\theta_i(t))))$$

Therefore we can pick

$$X_{\theta_i(0)}(f_i) = \psi_{\theta_i(0)}(A^{-1}_{\theta_i(0)}(f_i))$$

which is C^1 as composition of C^1 functions. We can perform an equivalent analysis for $\phi_j(0)$ and g_j to pick C^1 function $X_{\phi_j(0)}$. Let us now track the time derivative of $f_i(\theta_i)$ and $g_j(\phi_j)$

$$\dot{f}_i = \nabla_{\theta_i} f_i(\theta_i)\dot{\theta}_i = \|\nabla_{\theta_i} f_i(\theta_i)\|^2 \frac{\partial L}{\partial f_i}(F;G)$$

$$\dot{g}_j = \nabla_{\phi_j} g_j(\phi_j)\dot{\phi}_j = \|\nabla_{\phi_j} g_j(\phi_j)\|^2 \frac{\partial L}{\partial g_j}(F;G)$$

We can now replace $\theta_i = X_{\theta_i(0)}(f_i)$ and $\phi_j = X_{\phi_j(0)}(g_j)$ to get the equations required. □

> The XOR-functions in biological applications of Section 4 exemplify a set of cases where safety analysis is tractable. Much more generally, the separability assumption between the variables of f_i and g_j allows for a much more general positive result. Indeed, below we show that under standard assumptions for f_i and g_j and a target equilibrium $(p;q)$ with
>
> $$p_i \in (\max \mathrm{LocalMin}(f_i); \min \mathrm{LocalMax}(f_i))$$
> $$q_j \in (\max \mathrm{LocalMin}(g_j); \min \mathrm{LocalMax}(g_j))$$
>
> almost all initializations are safe by the Center-Stable Manifold-Theorem.

Theorem 2. If f_i and g_j have isolated stationary points, only strict saddle points, compact sublevel-sets, both equilibria $p_i \in (\max \mathrm{LocalMin}(f_i); \min \mathrm{LocalMax}(f_i))$ and $q_j \in (\max \mathrm{LocalMin}(g_j); \min \mathrm{LocalMax}(g_j))$, then almost all initializations are safe for $(p;q) \in \mathrm{Solution}(L)$.

Proof. Our proof structure goes as follows: Initially, we will prove that for each i and for all but a measure zero sets of $\theta_i(0)$ it holds that $p_i \in \mathrm{Im}_{f_i}(\theta_i(0))$ and $\theta_i(0)$ is not a stationary point of f_i. Correspondingly, for each j and for all but a measure zero sets of $\phi_j(0)$, it holds that $q_j \in \mathrm{Im}_{g_j}(\phi_j(0))$ and $\phi_j(0)$ is not a stationary points of g_j. Thus, since $(n;m)$ are finite, for all but a measure zero set of initializations $(\theta(0);\phi(0))$ are safe for $(p;q)$.

What remains to prove is that for almost all $\theta_i(0)$ it holds that $p_i \in \mathrm{Im}_{f_i}(\theta_i(0))$ and $\theta_i(0)$ is not a stationary point of f_i. The proof for g_j is completely symmetrical. Observe that the stationary points of f_i are isolated. So we clearly have that almost all $\theta_i(0) \in \mathbb{R}^{n_i}$ are not stationary points of f_i. We will then break the proof for $p_i \in \mathrm{Im}_{f_i}(\theta_i(0))$ in two pieces.

For the first piece we will prove that for almost all $\theta_i(0)$ we have that $p_i < \sup \mathrm{Im}_{f_i}(\theta_i(0))$. To do this we need to study the dynamics of θ_i as $t \to \infty$. Let $\theta_i(t)$ be the solution of $\dot{\theta}_i$. As $t \to \infty$, either $\|\theta_i(t)\|$ is bounded or it goes to ∞. For the case of $\|\theta_i(t)\| \to \infty$ we know that $f_i(\theta_i(t)) \to \infty$ since f_i has compact sublevel sets and thus $p_i < \sup \mathrm{Im}_{f_i}(\theta_i(0))$ follows directly. For the remaining cases, we know that $\theta_i(t)$ is bounded and thus it has a connected limit set. But since only stationary points of f_i can be limit sets of $\theta_i(t)$ and they are isolated, this means that $\theta_i(t)$ has exactly one limit point and as a result it converges to this limit point as $t \to \infty$. Let us call this point r. Since $f_i(\theta_i(t))$ is increasing because of the gradient ascent dynamics of $\dot{\theta}_i$, we have that $\sup \mathrm{Im}_{f_i}(\theta_i(0)) = f_i(r)$. Since r is a stationary point of f_i, it is either either a local minimum, a local maximum or a saddle point of f_i. Clearly r cannot be a local minimum since $f_i(\theta_i(t))$ is increasing so it cannot converge to a local minimum of f_i. Regarding saddle points, by assumption they are all strict and thus by the Stable Manifold Theorem (Theorem 12) only a zero measure set can converge to any of the isolated and thus countable saddle points. This leaves us with the r case of being a local maximum in which case we have that $f_i(r) \geq \min \mathrm{LocalMax}(f_i))$. But by assumption $p_i < \min \mathrm{LocalMax}(f_i)) \leq f_i(r) = \sup \mathrm{Im}_{f_i}(\theta_i(0))$.

For the second piece we will prove that for almost all $\theta_i(0)$ we have that $p_i > \inf \mathrm{Im}_{f_i}(\theta_i(0))$. To do this we need to study the dynamics of Σ_1 as $t \to \infty$. Let $\theta_i(t)$ be the solution of Σ_1. As $t \to \infty$, either $\|\theta_i(t)\|$ is bounded or it goes to ∞. For the case of $\|\theta_i(t)\| \to \infty$ we know that $f_i(\theta_i(t)) \to \infty$ since f_i has compact sublevel sets. But this is a contradiction since $f_i(\theta_i(t))$ is increasing so it cannot approach ∞ as $t \to \infty$. For the remaining cases, just like above we have that $\theta_i(t)$ converges to a point r and because of the gradient ascent dynamics of Σ_1 we have that $\inf \mathrm{Im}_{f_i}(\theta_i(0)) = f_i(r)$. Since r is a stationary point of f_i, it is either either a local minimum, a local maximum or a saddle point of f_i. Clearly r cannot be a local maximum since $f_i(\theta_i(t))$ is increasing so it cannot converge to a local maximum of f_i as $t \to \infty$. Once again for saddle points, we know from (Theorem 12) only a zero measure set can converge to any of the isolated and thus countable saddle points. This leaves us with the case of r being a local minimum in which case we have that $f_i(r) \le \max \mathrm{LocalMin}(f_i))$. But by assumption $p_i > \max \mathrm{LocalMin}(f_i)) \ge f_i(r) = \inf \mathrm{Im}_{f_i}(\theta_i(0))$. □

Remark 4. The key message of the above theorem is that the safety condition is a reasonable and relatively minimal assumption. Going beyond this type of result remains difficult as even in non-convex optimization, understanding which initializations lead to which local minimum of a loss function under gradient descent is a rather hard problem in its full generality, so no meaningful progress is possible without constraining ourselves in special cases with additional structure. Thus, instead of reiterating negative results based on unfortunate initializations that are not in agreement with the empirical success of GANs, we choose to study the dynamics of HCC games under safety.

C Hidden Convex Concave Games

In this section, we analyze the derived stability properties of the hidden convex concave games. It is worth mentioning that without strict/strong convexity/concavity from at least one of the operators, the quality of the results are limited to "Lyapunov Stability". Firstly, we present a construction of a Lyapunov function for the operators' dynamics Theorem 3. Then, in Theorem 4 and Theorem 5 we explore the stability of the initial conditions in the parameter space.

C.1 General case

The following theorem presents the construction of a Lyapunov potential function for the induced operator dynamics. To motivate its construction, we can study a fundamental convex-concave function $L(x;y) = (x-p)^2 - (y-q)^2$ with saddle point $(p;q)$. Under the gradient-descent-ascent dynamics

$$(T) := \begin{cases} \dot{x} = -\nabla_x L(x;y) & \text{(minimization of convex part)} \\ \dot{y} = \nabla_y L(x;y) & \text{(maximization of concave part)} \end{cases}$$

it is easy to check that $H(x;y) = (x-p)^2 + (y-q)^2$ meets all the criteria of a Lyapunov function. The construction below extends this argument to any convex-concave function $L(F;G)$ and bypasses the more complex multiplicative terms for the gradient induced dynamics of Theorem 1. Notice that

$$H(F;G) = \sum_{i=1}^{N} \int_{p_i}^{f_i} \frac{z-p_i}{\|\nabla f_i(X_{\theta_i(0)}(z))\|^2} dz + \sum_{j=1}^{M} \int_{q_j}^{g_j} \frac{z-q_j}{\|\nabla g_j(X_{\phi_j(0)}(z))\|^2} dz$$

coincides with the ℓ_2^2 distance from $(p;q)$ in the case of gradient norms equal to one, i.e.

$$\|\nabla f_i\|^2 = \|\nabla g_j\|^2 = 1$$

Lemma 2. If L is convex concave and $(\theta(0);\phi(0))$ is a safe for $(p;q)\in\text{Solution}(L)$, then the following quantity is non-increasing under the dynamics of Equation (3):

$$H(F;G)=\sum_{i=1}^{N}\int_{p_i}^{f_i}\frac{z-p_i}{\|\nabla f_i(\theta_{i(0)}(z))\|^2}dz+\sum_{j=1}^{M}\int_{q_j}^{g_j}\frac{z-q_j}{\|\nabla g_j(\phi_{j(0)}(z))\|^2}dz \tag{4}$$

Proof. Simple substitution gets us the following

$$\begin{aligned}\dot H &= \sum_{i=1}^{N}-(f_i-p_i)\frac{\partial L}{\partial f_i}(F;G)+\sum_{j=1}^{M}(g_j-q_j)\frac{\partial L}{\partial g_j}(F;G)\\ &= -\langle F-p;\nabla_F L(F;G)\rangle+\langle G-q;\nabla_G L(F;G)\rangle\end{aligned}$$

By Theorem 13 for the convex $L(\cdot;G)$ and concave $L(F;\cdot)$.

$$\begin{aligned}-\langle F-p;\nabla_F L(F;G)\rangle &\le L(p;G)-L(F;G)\\ \langle G-q;\nabla_G L(F;G)\rangle &\le L(F;G)-L(F;q)\end{aligned}$$

Thus we can end up writing

$$\begin{aligned}\dot H &\le L(p;G)-L(F;G)+L(F;G)-L(F;q)\\ &\le L(p;G)-L(p;q)+L(p;q)-L(F;q)\le 0\end{aligned}$$

The last inequality holds since $(p;q)\in\text{Solution}(L)$. Indeed, if $(p;q)$ is a saddle point of L then $L(p;G)\le L(p;q)\le L(F;q)$. □

Theorem 3. If L is convex concave and $(\theta(0);\phi(0))$ is a safe for $(p;q)\in\text{Solution}(L)$, then $(p;q)$ is stable for Equation (3).

Proof. Leveraging Lemma 2, there is a function H which is well defined in $D=\{\text{Im}_{f_i}(\theta_i(0))\}_{i=1}^{N}\times\{\text{Im}_{g_j}(\phi_j(0))\}_{j=1}^{M}$ and in this domain $\dot H\le 0$. Given the safety conditions we know that $(p;q)\in D$. Observe that for the proposed function, it holds that $H(p;q)=0$. Also for each f_i and g_j term in H we know that it has its minimum of value 0 at the corresponding p_i and q_j. We can deduce this by taking the derivative of each term to study its monotonicity. For example, the f_i terms are strictly increasing in $f_i>p_i$ and strictly decreasing in $f_i<p_i$. Thus for all $D-\{(p;q)\}$, $H>0$. Applying Theorem 9 for the continuously differentiable H we have that $(p;q)$ is stable for Equation (3). □

In the following example, we examine how it is possible to transfer the stability properties between two (topological conjugate) dynamical systems.

Theorem 4. If f_i and g_j are sigmoid functions and L is convex concave and there is a $(\theta(0);\phi(0))$ that is safe for $(p;q)\in\text{Solution}(L)$, then $(F^{-1}(p);G^{-1}(q))$ is stable for Equation (1).

Proof. Firstly, we recall the property of sigmoid's gradient:

$$\frac{d\sigma(x)}{dx}=\sigma(x)(1-\sigma(x)):$$

Thus the transformed dynamical system in the operator space can be written as:

$$(T):=\left\{\begin{aligned}\dot f_i &= -f_i^2(1-f_i)^2\tfrac{\partial L}{\partial f_i}(F;G)\\ \dot g_j &= g_j^2(1-g_j)^2\tfrac{\partial L}{\partial g_j}(F;G)\end{aligned}\right\}$$

Notice that

1. The dynamical system (T) in the operator space is <u>independent</u> of the initial conditions. In fact, the dynamical system of (T) and the one of Equation (1), called ( ) for short, are diffeomorphic for all initializations, not just a specific trajectory.
2. Since $(\theta(0);\phi(0))$ is safe, using Theorem 3 we get that $(p;q)$ is stable for (T).

We would like to prove that for every open neighborhood V of $(F^{-1}(p); G^{-1}(q))$ there exists an open neighborhood U of $(F^{-1}(p); G^{-1}(q))$ such that

$$(\ _{init};\ _{init}) \in U \Rightarrow \forall t \geq 0 : (\ (t);\ (t)) \in V:$$

Using the diffeomorphism $= \ _{!\ T}$ between GDA dynamics of (f) and (T), (V) is an open neighborhood of $(p; q)$ since V is open and $((F^{-1}(p); G^{-1}(q))) \quad (p; q) \in (V)$. By Item 2, since $(p; q)$ is stable for (T) there is an open neighborhood $\tilde{U}$ of $(p; q)$ such that:

$$(F_{init}; G_{init}) \in \tilde{U} \Rightarrow \forall t \geq 0 : (F(t); G(t)) \in (V)$$

or equivalently

$$(\ _{init};\ _{init}) \in \tilde{U} \Rightarrow \forall t \geq 0 : (\ (t);\ (t)) \in (V)$$

Indeed, using the inverse diffeomorphism[1], we can establish that for $U = \ ^{-1}(\tilde{U})$ it holds that

$$(\ _{init};\ _{init}) \in U \Rightarrow \forall t \geq 0 : (\ (t);\ (t)) \in V$$

□

Until now, we have established the stability of a pair $(p; q)$ for the induced dynamics (T). By the construction of the induced dynamics, (T) is coupled only with a very specific initial condition $(\ _{init};\ _{init})$. In order to tackle the challenge of a stability result for a whole region of initial conditions, in the following lemma we prove that $r(t;\ ) = \|F(\ ) - p\|^2 + \|G(\ ) - q\|^2$ can work like an intrinsic measure of closeness for the $\{\ ;\ \}$-parameter space around a hidden fixed point of the $\{F; G\}$-operator space. Under this "hidden" neighborhood notion, stability property can be taken by assuming the properness of the hidden operators.

Theorem 5. Let $(p; q) \in \mathrm{Solution}(L)$. Let R_{f_i} and R_{g_j} be the set of regular values[3] of f_i and g_j respectively. Assume that there is a $\ > 0$ such that $[p_i - \ ; p_i + \ ] \subseteq R_{f_i}$ and $[q_j - \ ; q_j + \ ] \subseteq R_{g_j}$. Define

$$r(t) = \|F(\ (t)) - p\|^2 + \|G(\ (t)) - q\|^2:$$

If f_i and g_j are proper functions[4], then for every $\ > 0$, there is an $\ > 0$ such that

$$r(0) < \ \Rightarrow \forall t \geq 0 : r(t) < \ :$$

Proof. Let us define the following sets

$$\forall i \in [n] \quad : A_i = \{\ _i \in \mathbb{R}^{n_i} \mid f_i(\ _i) \in [p_i - \ ; p_i + \ ]\}$$
$$\forall j \in [m] \quad : B_j = \{\ _j \in \mathbb{R}^{m_j} \mid g_j(\ _j) \in [q_j - \ ; q_j + \ ]\}$$

Since f_i and g_j are proper A_i and B_j are compact sets. Thus, the continuous functions $\|r f_i(\ _i)\|^2$ and $\|r g_j(\ _j)\|^2$ have a minimum and maximum value on A_i and B_j respectively. Let us call K_{f_i} and K_{g_j} the maxima and $\ _{f_i}$ and $\ _{g_j}$ the minima. Observe that the minima and maxima must be all greater than zero since $[p_i - \ ; p_i + \ ]$ and $[q_j - \ ; q_j + \ ]$ are regular values. Let us define

$$= \min\{\min_{1 \leq i \leq n} \ _{f_i};\ \min_{1 \leq j \leq m} \ _{g_j}\}$$
$$K = \max\{\max_{1 \leq i \leq n} K_{f_i};\ \max_{1 \leq j \leq m} K_{g_j}\}$$

where $K \geq \ > 0$ as we discussed. Let us create the following set

$$S = \{(\ ;\ ) \in \mathbb{R}^N \times \mathbb{R}^M \mid \forall i \in [n] : \ _i \in A_i;\ \forall j \in [m] : \ _j \in B_j\}$$

We can prove that every $(\ ;\ ) \in S$ is a safe initialization for $(p; q)$. Of course, every $\ _i$ and $\ _j$ are not stationary points of f_i and g_j respectively. We also need to prove that the equilibrium $(p; q)$ is feasible. We will prove this by contradiction. Let there be $(e; a) \in S$ such that $(p; q)$ is not feasible. Without loss of generality we can assume that there is $i \in [n]$ such that $p_i \notin \mathrm{Im}_{f_i}(\ _i)$.

---

[3] A value $a \in \mathrm{Im}\, f$ is called a regular value of f if $\forall q \in \mathrm{dom}\, f : f(q) = a$, it holds $r f(q) \neq 0$.

[4] A function is proper if inverse images of compact subsets are compact.

The case for the g_j is symmetrical. Along the gradient ascent trajectory of f_i with initialization at $_i$, observe that $f_i(t)$ cannot attain an infimum or a supremum in $[p_i \;\; ; p_i + \;\;]$ because there are no stationary points of f_i in A_i. Observe also that at initialization $f_i(\;_i) \in [p_i \;\; ; p_i + \;\;]$. Thus $[p_i \;\; ; p_i + \;\;] \;\; \mathrm{Im}_{f_i}(\;_i)$, a contradiction.

Let us pick an initialization $(\;(0); \;(0))$ such that $r(0) \;\; ^2$. It is clear that $(\;(0); \;(0)) \in S$ and so it is safe for $(p; q)$. We can do the same steps as in Theorem 3 to prove that the function $H(F; G)$ below does not increase under the dynamics of Equation (1):

$$H(F; G) = \sum_{i=1}^{N} \int_{p_i}^{f_i} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz + \sum_{j=1}^{M} \int_{q_j}^{g_j} \frac{z \;\; q_j}{\| \nabla g_j(X_{j(0)}(z)) \|^2} dz$$

Observe that since $(\;(0); \;(0)) \in S$ we have that the interval between p_i and $f_i(\;_i(0))$ belongs in $[p_i \;\; ; p_i + \;\;]$ and $\| \nabla f_i(\;) \|^2 \;\; $ in this interval. Thus we can write

$$\frac{(f_i(\;_i(0)) \;\; p_i)^2}{2} \;\; \int_{p_i}^{f_i(\;_i(0))} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz$$

Repeating the same argument for all f_i and g_j we have that

$$\frac{r(0)}{2} \;\; H(F(\;(0)); G(\;(0))) \;\; H(F(\;(t)); G(\;(t)))$$

Let us pick $r(0) < \min\{\;^2; \;^2 \frac{}{K}\} = \;^2 \frac{}{K}$. We already know that trajectories start in S. We will prove that they also remain in S. We will do this by contradiction. If a trajectory escapes S, then without loss of generality this means that there is at least an $i \in [n]$ such that at some $t > 0$, $f_i(\;_i(t)) \notin [p_i \;\; ; p_i + \;\;]$. The case of g_j is similar. Clearly we have that

$$\int_{p_i}^{f_i(\;_i(t))} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz \;\; \min\left\{ \int_{p_i}^{p_i \;\;} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz; \int_{p_i}^{p_i +} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz \right\}$$

As above, we have that the gradients in the integrals of the right hand side are less or equal than K so

$$\int_{p_i}^{f_i(\;_i(t))} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz \;\; \frac{^2}{2K}:$$

The terms of H are all non-negative so we have that

$$\frac{r(0)}{2} \;\; H(F(\;(t)); G(\;(t))) \;\; \int_{p_i}^{f_i(\;_i(t))} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz \;\; \frac{^2}{2K}:$$

But $r(0) < \;^2 \frac{}{K}$, a contradiction. So the trajectories will stay in S. We can then write

$$\int_{p_i}^{f_i(\;_i(t))} \frac{z \;\; p_i}{\| \nabla f_i(X_{i(0)}(z)) \|^2} dz \;\; \frac{(f_i(\;_i(t)) \;\; p_i)^2}{2K}:$$

Repeating the same argument for all f_i and g_j we have that

$$\frac{r(0)}{2} \;\; H(F(\;(t)); G(\;(t))) \;\; \frac{r(t)}{2K}:$$

For every $\;> 0$, there is a positive $\;= \frac{\min\{\;^2; \;\}}{K}$ such that

$$r(0) < \;\; \Longrightarrow \;\; r(t) < \;:$$

□

A special case of the above result is the standard convex-concave games:

Corollary 3. *Let $L(x; y)$ be strictly convex concave and $\mathrm{Solution}(L)$ is the non empty set of equilbria of L. Then $\mathrm{Solution}(L)$ is locally asymptotically stable for continuous GDA dynamics.*

Proof. The proof of the above classical result can be derived by the straightforward application of Lemma 3 for the case of $F(x) = x$ and $G(y) = y$. Notice that i) if $F; G$ are the identity maps all the initial configurations are safe and ii) if $\| \nabla F\|^2 = \| \nabla G\|^2 = 1$, then the initialization-dependent Lyapunov functions coincide to a single Lyapunov function, which is actually the squared Euclidean distance $r(\;;\;) = \|F(\;) \;\; p\|^2 + \|G(\;) \;\; q\|^2 = \| \;\; p\|^2 + \| \;\; q\|^2$. □

C.2 Hidden strictly convex concave games

C.2.1 Gradient Descent-Ascent Dynamics

> In the following preliminary result, we show that strict convexity or concavity of $L(\cdot,\cdot)$, for at least one of its arguments, suffices to yield locally asymptotic stability starting from a safe initial condition. Our argumentation leverages the power of Theorem 10 and combines the previous section stability results. Here, we will firstly outline the basic steps below:
>
> 1. We start by showing that there exists a compact set $\Omega \subset D$.
> 2. Therefore, since $\dot{H} \le 0$ (Lyapunov property), any configuration $(F(0), G(0))$ starting from a bounded sub-level set Ω of H, will remain inside Ω over all time.
> 3. The second crucial observation is that thanks to the strictness on convexity or concavity of L, the largest invariant set of $\dot{H} = 0$ contains only points belonging to Von Neumann's Solution(L).
>
> Then Theorem 10 implies the local asymptotic stability of set Z for Equation (3).

Lemma 3. Let L be strictly convex concave and $Z \subseteq$ Solution(L) is the non empty set of equilbria of L for which $(\theta(0), \phi(0))$ is safe. Then Z is locally asymptotically stable for Equation (3).

Proof. Pick a point $(p, q) \in Z$. Since our initialization is safe for this saddle point, we can construct the H function as in Theorem 3 and prove that it has the following property

$$\dot{H} \le 0 \text{ in } D = \{\mathrm{Im}_{f_i}(\Theta_i(0))\}_{i=1}^{N} \times \{\mathrm{Im}_{g_j}(\Phi_j(0))\}_{j=1}^{M}$$

If $(F(\theta(0)), G(\phi(0))) = (p, q)$ then the theorem holds trivially. Otherwise, take a ball B centered at the equilibirum with a small enough radius such that it is contained in the interior of D

$$H_0 = \min_{(F,G) \in \partial B} H(F, G)$$
$$\Omega = \{(F, G) \in B \,|\, H(F, G) \le H_0/2\}$$

We know that in both of the cases $H_0 > 0$ from Theorem 3.

Since $\dot{H} \le 0$, starting in Ω, it implies that $H(F(t), G(t)) \le H_0$ for $t \ge 0$, so Ω is forward invariant. Since $\Omega \subset D$ we know that it is bounded. Ω is closed since it is a sublevel set of a continuous function. Notice that the restriction of Ω on B does not affect the above properties since Ω is in the interior of B. Thus Ω is a compact forward invariant set, satisfying the requirement of Theorem 10

Let $E = \{(F, G) \in B \,|\, \dot{H}(F, G) = 0\}$. Without loss of generality we can assume that $L(\cdot, q)$ is strictly convex as the case of $L(p, \cdot)$ being strictly concave is similar. In the following inequality

$$\dot{H} \le L(p, G) - L(p, q) + L(p, q) - L(F, q) \le 0$$

we know that $L(p, G) - L(p, q) \le 0$ and $L(p, q) - L(F, q) \le 0$.

So $\dot{H} = 0$ implies $L(p, G) = L(p, q) = L(F, q)$. By the strict convexity of $L(\cdot, q)$ we know that this means that $F = p$. Let M be the largest invariant set inside E. By the properties of M being invariant subset of E we have

$$(F(0), G(0)) \in M \implies \forall t : F(t) = p \text{ and } L(p, G(t)) = L(p, q)$$

Taking the time derivatives on each of the constant quantities, they should be zero.

$$\dot{f}_i = 0 \implies \quad \forall i \in [N] : \quad \|\nabla_{\theta_i} f_i(X_{\theta_i(0)}(p_i))\|^2 \frac{\partial L}{\partial f_i}(p, G) = 0$$

$$\dot{L}(p, G(t)) = 0 \implies \quad \sum_{j=1}^{M} \|\nabla_{\phi_j} g_j(X_{\phi_j(0)}(g_j))\|^2 \left(\frac{\partial L}{\partial g_j}(p, G)\right)^2 = 0$$

We know that $\|\nabla_i f_i(X_{i(0)}(p_i))\| \neq 0$ by the safety conditions and that $\|\nabla_j g_j(X_{j(0)}(g_j))\|^2 \neq 0$ inside D again by safety conditions. This implies

$$\forall i \in [N]: \frac{\partial L}{\partial f_i}(p;G) = 0$$
$$\forall j \in [M]: \frac{\partial L}{\partial g_j}(p;G) = 0$$

Thus M contains only stationary points of L so $M \subseteq Solution(L)$. In addition $M \subseteq D$ so only stationary points of L for which the initialization is safe are allowed so $M \subseteq Z$. Applying Theorem 10 we have that for any initialization of Equation (3) inside Ω, as $t \to \infty$ $(F(t);G(t))$ approaches M and thus Z is locally asymptotically stable for Equation (3). □

A special case of the above result is the standard convex-concave games:

Corollary 4. Let $L(x;y)$ be strictly convex concave and $Solution(L)$ is the non empty set of equilbria of L. Then $Solution(L)$ is locally asymptotically stable for continuous GDA dynamics.

In the following main result of our work, we show that strict convexity or concavity in $L(\cdot,\cdot)$, for at least one of its arguments, suffices to yield a convergence result to a Von Neumann's $Solution(L)$ starting from a safe initial condition. In order to get convergence results for any safe initialization, we need to study the region of attraction of the set $Z \subseteq Solution(L)$. We refine the estimation of the region of attraction as proposed in Lemma 3 by analyzing the behavior of the level sets of H. More precisely, we show that the proposed Lyapunov function

$$H(F;G) = \sum_{i=1}^{N} \int_{p_i}^{f_i} \frac{z - p_i}{\|\nabla f_i(X_{i(0)}(z))\|^2} dz + \sum_{j=1}^{M} \int_{q_j}^{g_j} \frac{z - q_j}{\|\nabla g_j(X_{j(0)}(z))\|^2} dz$$

is radially unbounded. In other words, while the operators converges to their limit values (supremum/infimum of their domain) $H \to +\infty$. In order to show that we analyze the asymptotic behavior of $\int_c^F \frac{1}{\|\nabla f_i\|^2}$, while $F \to \sup f_i$. Hence,

A) Theorem 10 implies that the trajectory will approach the set of stationary points of H or equivalently a set of Von Neumann's $Solution(L)$.

B) The stability of $Solution(L)$ and Theorem 11, leads to the conclusion that the trajectory will converges to a specific point of $Solution(L)$.

Theorem 6. Let L be strictly convex concave and $Z \subseteq Solution(L)$ is the non empty set of equilbria of L for which $(\theta(0);\phi(0))$ is safe. Under the dynamics of Equation (1) $(F(\theta(t));G(\phi(t)))$ converges to a point in Z as $t \to \infty$.

Proof. Again let's pick a point $(p;q) \in Z$. Since our initialization is safe for this saddle point, we can construct the H function as in Theorem 3 and prove that it has the following property

$$\dot{H} \leq 0 \text{ in } D = \{Im_{f_i}(\theta_i(0))\}_{i=1}^{N} \times \{Im_{g_j}(\phi_j(0))\}_{j=1}^{M}$$

If $(F(\theta(0));G(\phi(0))) = (p;q)$ then the theorem holds trivially. Otherwise define

$$H_0 = H(F(\theta(0));G(\phi(0)))$$
$$\Omega = \{(F;G) \in D \mid H(F;G) \leq H_0\}$$

where we know that $H_0 > 0$ from Theorem 3. Let us assume that indeed Ω is in the interior of D. Then, applying the same argumentation as in Lemma 3 combined with Theorem 3, all fixed points in Z are stable. So applying Theorem 11 we get that the trajectory initialized at $(F(\theta(0));G(\phi(0))) \in \Omega$ converges to a point in Z. It remains to prove our assertion about the set Ω.

Claim 1. Ω is in the interior of D.

Proof. We will argue that as $(F;G)$ approaches the boundary of D, the value of H should become unbounded. If this is true then for the finite upper bound of H_0, should have no points close to the boundary of H and thus it should be in the interior.

As $(F;G)$ approach the boundary of D, at least one of the variables f_i or g_j approaches the endpoints points of $\mathrm{Im}_{f_i}(\ _i(0))$ or $\mathrm{Im}_{g_j}(\ _j(0))$ respectively. We will study the case of f_i since the case of g_j is symmetrical. The endpoint f_{is} can be either the supremum or the infimum of the gradient ascent trajectory on f_i or ∞ if they do not exist. Let f_{is} be the supremum or ∞ depending on if the former exists. We can take the gradient ascent dynamics and apply Lemma 1 to get

$$\dot{f}_i = \|\nabla_{\ i} f_i(X_{\ i(0)}(f_i))\|^2$$

We know that $f_i(\ _i(t))$ goes to f_{is} when initialized at $f_i(\ _i(0))$. Let us define the following function

$$a(f_i) = \int_{p_i}^{f_i} \frac{1}{\|\nabla f_i(X_{\ i(0)}(z))\|^2} dz$$

Observe that $\dot{a} = 1$, thus $\lim_{t\to\infty} a(f_i(t)) = \infty$. In other words

$$\lim_{t\to\infty} \int_{p_i}^{f_i(t)} \frac{1}{\|\nabla f_i(X_{\ i(0)}(z))\|^2} dz = \int_{p_i}^{f_{is}} \frac{1}{\|\nabla f_i(X_{\ i(0)}(z))\|^2} dz = \infty$$

Symmetrically if f_{is} is the infimum or $-\infty$, then the limit above would be $-\infty$. In either case

$$f_i \to f_{is} \implies \int_{p_i}^{f_i} \frac{z - p_i}{\|\nabla f_i(X_{\ i(0)}(z))\|^2} dz \to \infty$$

For the last step it is important to note that p_i is not at the boundary of D based on the safety conditions. Therefore as $(F;G)$ approach the boundary of D in the dynamics of Equation (3), at least one of the terms of H goes to infinity. Also note that all the terms of H are individually non-negative so no matter what the other variables in $(F;G)$ are doing they cannot stop $H \to \infty$. □

□

Again, a special case of the above result is the standard convex-concave games:

Corollary 5. *Let $L(\mathbf{x};\mathbf{y})$ be strictly convex concave and $\mathrm{Solution}(L)$ is the non empty set of equilbria of L. Under the continuous GDA dynamics $(\mathbf{x}(t);\mathbf{y}(t))$ converges to a point in $\mathrm{Solution}(L)$ as $t \to \infty$.*

C.2.2 Connections to Hamiltonian Descent

In GANs numerous learning heuristics are being tested and explored. One technique that has particular interesting theoretical justification as well as practical performance is Hamiltonian Gradient Descent (HGD). Understanding the convergence guarantees for HGD is an open research question [39, 8, 51]. We provide some new justification about its success in GANs by provably establishing convergence of a modified version of HGD in a relatively simple but illustrative subclass of hidden convex concave games, namely 2x2 hidden bi-linear games. This class of games is fairly expressive. Despite the restriction of planar bi-linear competition in the output space, the hidden game can have an arbitrary number of variables in the parameter space. It's important to note that given the bi-linear nature of competition, the classical GDA dynamics cycles instead of converging to the equilibrium as shown in [65]

More precisely, in the hidden 2x2 bi-linear game presented in [65], we have two functions $f : \mathbb{R}^N \to [0;1]$ and $g : \mathbb{R}^M \to [0;1]$ and two constants $(p;q) \in (0;1)^2$ where $(p;q)$ is the fully mixed equilibrium of the bi-linear game. Without loss of generality, we are interested in solving the following problem

$$\min_{\ \in \mathbb{R}^M} \max_{\ \in \mathbb{R}^N} (f(\ ) - p)(g(\ ) - q)$$

Defining $L(\ ;\ ) = (f(\ ) - p)(g(\ ) - q)$, the dynamics of HGD are:

$$\begin{aligned} \dot{} &= -\frac{1}{2}\nabla \|\nabla L(\ ;\ )\|^2 - \frac{1}{2}\nabla \|\nabla L(\ ;\ )\|^2 \\ \dot{} &= -\frac{1}{2}\nabla \|\nabla L(\ ;\ )\|^2 - \frac{1}{2}\nabla \|\nabla L(\ ;\ )\|^2 \end{aligned} \tag{6}$$

Observe that the second term of each right hand side would be zero in a classical bi-linear game but involves second order derivatives of f and g in the case of hidden bi-linear games. To circumvent the complexities of the second order derivatives and mimic the classical bi-linear game we will study a modified version of Equation (6), namely:

$$\_ = \frac{1}{2} r \quad kr \quad L(;)k^2 \qquad \_ = \frac{1}{2} r \quad kr \quad L(;)k^2 \tag{7}$$

Employing an analysis similar to the one in Section 3.2, we get the following convergence result:

Theorem 15. Let $((0); (0))$ be safe for $(p; q)$. Then $(f((t)); g((t)))$ converges to $(p; q)$ under the dynamics of Equation (7).

Proof. Simple substitution gives us

$$\_ = r \quad f( )kr \quad g( )k^2 (f( ) \quad p)$$
$$\_ = r \quad g( )kr \quad f( )k^2 (g( ) \quad q)$$

Applying Lemma 1 and following the same steps as before

$$f\_ = kr \quad f(X_{(0)}(f))k^2 kr \quad g(X_{(0)}(g))k^2 (f \quad p)$$
$$g\_ = kr \quad g(X_{(0)}(g))k^2 kr \quad f(X_{(0)}(f))k^2 (g \quad q)$$

Once again we consider the function

$$H(f; g) = \int_p^f \frac{z \quad p}{kr \; f(X_{(0)}(z))k^2} dz + \int_q^g \frac{z \quad q}{kr \; g(X_{(0)}(z))k^2} dz$$

Simple substitution gives

$$H\_ = (f \quad p) \; kr \quad g(X_{(0)}(g))k^2 (f \quad p) \qquad (g \quad q) \; kr \quad f(X_{(0)}(f))k^2 (g \quad q)$$

A little bit of reorganization gives

$$H\_ = (f \quad p)^2 kr \quad g(X_{(0)}(g))k^2 \quad (g \quad q)^2 kr \quad f(X_{(0)}(f))k^2 \quad 0$$

Thus, we get

$$H\_ \quad 0 \text{ in } D = \mathrm{Im}_f((0)) \quad \mathrm{Im}_g((0))$$

Similarly with the strict convex analysis of the previous section, if $(f((0)); g((0))) = (p; q)$ then the theorem holds trivially. Otherwise define

$$H_0 = H(f((0)); g((0)))$$
$$= f(f; g) \; 2 \; D j H(f; g) \quad H_0 g$$

where we know that $H_0 > 0$ from Theorem 3. Additionally, we can apply Claim 1 even in the new dynamics, so is in the interior of D. Since $H\_ \quad 0$, starting in , it implies that $H(f(t); g(t)) \quad H_0$ for $t \quad 0$, so $f(t); g(t)$ stays in . Additionally, is closed since it is a sublevel set of a continuous function. Notice that the restriction of on D does not affect the above properties since is in the interior of D. Thus is a compact forward invariant set.

For a safe initialization $((0); (0))$, both $kr \quad g(X_{(0)}(g(t)))k; kr \quad f(X_{(0)}(f(t)))k$ cannot go to 0 as this happens only at the boundaries of D which are outside . So $H\_ = 0$ only at $(p; q)$ in .

Therefore, applying Theorem 10, we get that $(f((t)); g((t)))$ converges to $(p; q)$

□

C.3 Regularization and convergence

In this section, we show that even in the absence of strict convexity/concavity for both of the operators, it is possible to achieve a positive convergence result by sacrificing the exactness of a targeted equilibrium. In other words, we prove that by adding a small regularization term, the new utility function becomes strictly convex strictly concave. Beside the guaranteed convergence of the "perturbed" L^0, we can always choose sufficiently small magnitude of regularization such that the new equilibria are arbitrarily close to the initial ones.

Theorem 7. If L is a convex concave function with invertible Hessians at all its equilibria, then for each > 0 there is a > 0 such that L^0 has equilibria that are -close to the ones of L.

Proof. For any choice of > 0 we have that L^0 is strictly convex strictly concave so the KKT conditions are sufficient to determine its equilibria.

$$\frac{@L(x;y)}{@x_i} + x_i = 0$$

$$\frac{@L(x;y)}{@y_j} \quad y_j = 0$$

We can view the above set of constraints as a single vector constraint $r(; x; y) = 0$. Note that by assumption of the Hessians being invertible at all equilibria, L has a unique equilibrium $(x ; y)$. Clearly we have that $r(0; x ; y) = 0$. Observe that for the Jacobian of r at $(0; x ; y)$ with respect to (x, y) we have that

$$D_{(x;y)} r(0; x ; y) = r^2 L(x ; y)$$

and thus it is invertible. Invoking the Implicit function Theorem, there is a differentiable function g, defined in a small enough neighborhood of 0, that takes a and returns $g( ) = (x( ); y( ))$ such that $r(; g( )) = 0$. Thus for a small enough , we have that g returns the corresponding equilibria of L^0. By continuity of g, for all there is a > 0

$$8\,0 < < \; : \|x( ) \quad x(0)\|^2 + \|y( ) \quad y(0)\|^2 \quad {}^2$$

But $(x(0); y(0)) = (x ; y)$ so the equilbrium of L^0 has an -close equilibrium of L for $<$. By strict convexity strict concavity of L^0, it has a unique equilibrium as well. So the equilibria of L^0 and L are -close to each other. □

The previous theorem highlights that small values of induce only small changes to the equilibria of the hidden game. As is the case for classical convex concave games, larger values of lead to (exponentially) faster convergence. To prove this for HCC games, we provide a detailed upper and lower bound analysis of the gradients of f_i and g_j.

Theorem 8. Let $((0); (0))$ be a safe initialization for the unique equilibrium of $L^0(p; q)$. If

$$r(t) = \|F((t)) \quad p\|^2 + \|G((t)) \quad q\|^2$$

then there are initialization dependent constants $c_0, c_1 > 0$ such that $r(t) \quad c_0 \exp(\quad c_1 t)$.

Proof. Following the same analysis with the strict convex concave analysis of the previous section, if $(F((0)); G((0))) = (p; q)$ then the theorem holds trivially. Otherwise, since our initialization is safe for $(p; q)$, we can construct the H function as in Theorem 3 and prove that it has the following property in $D = f\, Im_{f_i}(_i(0))g_{i=1}^N \quad f\, Im_{g_j}(_j(0))g_{j=1}^M$

$$H\!\!L \quad L^0(p; G) \quad L^0(p; q) + L^0(p; q) \quad L^0(F; q)$$

$$\frac{}{2}\; \|F((t)) \quad p\|^2 + \|G((t)) \quad q\|^2$$

$$\frac{}{2} r(t)$$

Where the second step follows from $L^0(p; \ )$ being strongly concave and $L^0(\ ; q)$ being strongly convex and q, p being the corresponding optima of these functions since (p, q) is an equilibrium. Let us de ne

$$H_0 = H(F(\ (0)); G(\ (0)))$$
$$= \ f(F; G) \ 2 \ D j H(F; G) \quad H_0 g$$

where we know that $H_0 > 0$ from Theorem 3. Additionally, we can apply Claim 1 even in the new dynamics, so is in the interior of D. Since $\dot{H} \quad 0$, starting in , it implies that $H(F(\ (t)); G(\ (t))) \quad H_0$ for $t \quad 0$, so $(F(t); G(t))$ stays in . Additionally, is closed since it is a sublevel set of a continuous function. Notice that the restriction of D does not affect the above properties since is in the interior of D. Thus is a compact forward invariant set.

For a safe initialization $(\ (0); \ (0))$, the following continuous functions must have a minimum and maximum value on respectively.

$$K_{f_i} \quad kr \quad f_i(X_{\ i(0)}(\ ))k^2 \qquad {}_{f_i}$$
$$K_{g_j} \quad kr \quad g_j(X_{\ j(0)}(\ ))k^2 \qquad {}_{g_j}$$

Observe that the minima and maxima must be all greater than zero , since both $kr \ _j \ g_j(X_{\ j(0)}(g(t)))k; kr \ _i \ f_i(X_{\ i(0)}(f(t)))k$ cannot go to 0 as this happens only at the boundaries of D which are outside .

Let us de ne

$$= \min f \min_{1 \ i \ n} \ {}_{f_i}; \min_{1 \ j \ m} \ {}_{g_j} g$$
$$K = \max f \max_{1 \ i \ n} K_{f_i}; \max_{1 \ j \ m} K_{g_j} g$$

Observe that $K \quad kr \quad f_i(X_{\ i(0)}(\ ))k^2$ in this interval. Thus we can write

$$\frac{(f_i(\ _i(t)) \quad p_i)^2}{2} \qquad \int_{p_i}^{f_i(\ _i(t))} \frac{z \quad p_i}{kr \ f_i(X_{\ i(0)}(z))k^2} dz \qquad \frac{(f_i(\ _i(t)) \quad p_i)^2}{2K}$$

Repeating the same argument for all f_i and g_j we have that

$$\frac{r(t)}{2} \quad H(F(\ (t)); G(\ (t))) \quad \frac{r(t)}{2K}$$

Thus we can extend our analysis

$$\dot{H} \qquad r(t) \qquad \frac{2}{2} H(t) \ ) \quad H(t) \quad H_0 e^{\ t} \ ) \quad r(t) \quad 2 \quad K \quad H_0 e^{\ t}$$

□

D Applications

D.1 Connecting GANs and Hidden Convex-Concave Games

At the heart of many GAN formulations like the standard GAN [26], f-GAN [50] and Wassertein GAN (WGAN) [4] lies a classical convex concave game in the operator output space. Indeed for the realizeable case, [26] used the underlying convexity properties to find the Nash equilibria of standard GAN and [23] did the same thing for the f-GAN and WGAN. Perhaps surprisingly, neither work references explicitly the convex concave nature of the operator output space game or von Neumann's minimax theorem. To highlight the significance of von Neumann equilibria as a solution concept for GANs, we show how the optimal G and D can be derived separately from each other by solving the corresponding min-max (max-min) problems. This allows one to independently verify the validity of von Neumann's minimax theorem and its generalizations for GANs. We also extend our analysis to a wide class of non-realizeable cases as well.

In practice however, as noted explicitly by [27], the updates in GAN training happen in the parameter space giving rise to a HCC game. This has exactly motivated studying the learning dynamics of HCC games in Section 3.

Thus, in this section, we present these connections between Hidden Convex-Concave games and the different architectures of Generative Adversarial Networks. More specifically, we start by exploring the structure of GANs and we verify their hidden convex-concave intrinsic form.

1. Under this scope of hidden games, the strong (or even strict) convexity/concavity of at least one of the players (Discriminator/Generator) in combination with the convergence results of the following sections provide some theoretical explanation about the convergence properties of those architectures even under the vanilla Gradient Descent-Ascent Dynamics.

2. To indicate the relation of Von-Neumann solution with this hidden model, we leverage this hidden convex-concave structure in order to compute the well-known both min max and max min optima of GANs under the realizability or not assumption. The results of this section are summarized in the following table:

Type of GAN	G	D	Hidden Structure
GAN	p_{data}	$\frac{1}{2}$	Linear VS Strongly-Concave
GAN	$\arg\min_{G \in \mathcal{G}} JSD(p_{data} \| p_G)$	$\frac{p_{data}}{p_{data} + p_G}$	Linear VS Strongly-Concave
f-GAN	p_{data}	$f'(1)$	Linear VS Concave
f-GAN	$\arg\min_{G \in \mathcal{G}} D_f(p_{data} \| p_G)$	$f'\left(\frac{p_{data}}{p_G}\right)$	Linear VS Concave
WGAN	p_{data}	c	Linear VS Linear
WGAN	$\arg\min_{G \in \mathcal{G}} EMD(p_{data} \| p_G)$	–	Linear VS Linear

Table 1: p_{data} represents the target data distribution. G is the min-max generator and D is the max-min discriminator. JSD denotes the Jensen–Shannon divergence, D_f the f-divergence for the convex function f and EMD the earth mover distance and c the constant discriminator. xGAN, xGAN correspond to the realizable and the non-realizable case accordingly. – indicates the lack of a closed form solution for D of WGAN.

In the following three subsections, we analyze both the derivation of arg min max and arg max min for the "vanilla-GANs", f-GANs, W-GANs using min-max optimization arguments based on the Minimax Theorem for convex-concave functions. More precisely,

1. In the Lemmas 4, 9 and 14, we present the optimal discriminators which consist the best-response for the case of a fixed generator. In all these maximization problems, typically each $D(x)$ is decoupled and $D^*_G(x)$ is derived by the hidden concavity of the discriminator architecture.
2. In the Lemmas 5, 10 and 15, we present the optimal generators which consist the best-response for the case of a fixed discriminator. In all these minimization problems, typically the generator can cheat the fixed discriminator by producing greedily a distribution only over the restricted subset of the points for which the discriminator has the highest confidence about their originality.
3. In the Lemmas 6, 11 and 16, we leverage lemmas of (Item 1) to understand the form GAN's utility function which corresponds typically to JSD, f-divergence and Wasserstein distance which donate their name to their GAN architecture as well. Thus, it is then trivial to show that p_{data} is the optimal choice in the realizable case.
4. In the Lemmas 7, 12 and 17, on the other side of the coin, we emphasize to derive the minmax solutions too. Our proof strategy invokes the partition to two basic sets, S_{G_D} and $S^c_{G_D}$, the "preferable" or not data points by the generator. Leveraging the concavity part of the objective, we show that the best strategy for the discriminator is to label all the points uniformly with the same confidence in order to incentivize the generator to expands its support to the maximum possible.
5. In the Lemmas 8 and 13, we analyze the non-realizable case. One the one hand using Item 3 we are able to compute the arg max min generator G^*. To conclude about the arg min max discriminators we apply the Von Neumann's Minimax theorem to prove $D^* =$ Best-Response(G^*).

D.1.1 GAN

The utility of the zero-sum game $V(G; D)$ for the distribution p_{data} over the discrete set $\mathcal{N}$ is

$$V(G; D) = \sum_{x \in \mathcal{N}} p_{data}(x) \log(D(x)) + \sum_{x \in \mathcal{N}} p_G(x) \log(1 - D(x))$$

On the one hand, it is easy to check that for a fixed discriminator D, the utility function is linear over the p_G operator. On the other hand, for a fixed generator G, the utility function is of the form $a \log(D) + b \log(1 - D)$ which is strongly-concave.

We start our work with the following lemmas

Lemma 4 ([26]). For a fixed generator G the optimal discriminator is

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)}$$

Proof. Observe that the optimization problem for each $D(x)$ is decoupled. Thus

$$D^*_G(x) = \arg\max_{D \in [0;1]} p_{data}(x) \log(D) + p_G(x) \log(1 - D)$$

By concavity the unique maximum of the above is given by

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)}$$

□

Lemma 5. For a fixed discriminator D, any distribution supported only on

$$S_{G_D} = \{x \in \mathcal{N} : \forall x' \in \mathcal{N} \quad D(x) \le D(x')\}$$

is an optimal generator when it is allowed to choose any distribution over $\mathcal{N}$.

Proof. Observe that for a fixed discriminator, the optimal generator optimizes

$$\sum_{x \in N} p_G(x) \log(1 - D(x))$$

since the other term is independent of the generator. Let us define the following

$$D_{max} = \max_{x \in N} D(x)$$

Then we have that

$$\sum_{x \in N} p_G(x) \log(1 - D(x)) \geq \log(1 - D_{max})$$

with the equality being true only for distributions supported only on S_{G_D}. □

Lemma 6 ([26])**.** The min-max generator is the following distribution

$$G^* = \arg\min_{G \in \mathcal{G}} JSD(p_{data} || p_G).$$

Proof. We can substitute in $V(G; D)$ the optimal discriminator from Lemma 4. Thus we get

$$V(G; D^*_G) = \sum_{x \in N} p_{data}(x) \log \frac{p_{data}(x)}{p_{data}(x) + p_G(x)} + \sum_{x \in N} p_G(x) \log \frac{p_G(x)}{p_{data}(x) + p_G(x)}$$

We can now prove that

$$\begin{aligned} V(G; D^*_G) &= -\log(4) + KL\left(p_{data} \Big|\Big| \frac{p_G + p_{data}}{2}\right) + KL\left(p_G \Big|\Big| \frac{p_G + p_{data}}{2}\right) \\ &= -\log(4) + 2JSD(p_{data} || p_G) \end{aligned}$$

By minimizing $V(G; D^*_G)$, the result follows trivially. □

Lemma 7. The max-min discriminator is

$$\forall x \in N : D^*(x) = \frac{1}{2}$$

when the generator is allowed choose any distribution over N.

Proof. We can substitute in $V(G; D)$ the optimal generator from Lemma 5

$$\begin{aligned} V(G_D; D) = \log(D_{max}) &\sum_{x \in S_{G_D}} p_{data}(x) + \log(1 - D_{max}) \sum_{x \in S_{G_D}} p_G(x) \\ &+ \sum_{x \notin S_{G_D}} p_{data}(x) \log(D(x)) + \sum_{x \notin S_{G_D}} \underbrace{p_G(x)}_{0} \log(1 - D(x)) \end{aligned}$$

Let us define $D_{small} = \{D(x) | x \notin S_{G_D}\}$. Observe that if $|D_{small}| > 1$ then setting $D(x) = \max(D_{small})$ for each $x \notin S_{G_D}$ improves utility. Thus for the optimal discriminator we have $|D_{small}| = 1$. Let us call D_{min} the unique element of D_{small}. Then we have that

$$x \notin S_{G_D} \implies D^*(x) = D_{min}$$
$$x \in S_{G_D} \implies D^*(x) = D_{max}$$

Observe that for any combination of D_{max} and D_{min} with $D_{max} > D_{min}$, the constant discriminator D_{max} has higher utility. Therefore we can focus our attention on the constant discriminator $D_{const}(x) = D$

$$V(G_{D_{const}}; D_{const}) = \log(D) + \log(1 - D)$$

The optimal value for D is $\frac{1}{2}$ and as a result

$$D^*(x) = \frac{1}{2}.$$

□

Lemma 8 (Non-realizable case). If we assume that choice of generator G is restricted in $\mathcal{G}$, a convex compact subset of the $|N|$ dimensional simplex, such that $p_{data} \notin \mathcal{G}$. Then

$$(G^*, D^*) = \left(\arg\min_{G \in \mathcal{G}} JSD(p_{data} \| p_G), \frac{p_{data}}{p_{data} + p_{G^*}} \right)^5$$

Proof. We cannot readily apply von Neumann's minimax theorem since the $V(G, D)$ may be infinite at the boundary points of $\mathcal{D} = (0, 1)^{|N|}$ for the discriminator. We can still apply Fan's Minimax Theorem

$$\min_{G \in \mathcal{G}} \sup_{D \in \mathcal{D}} V(G, D) = \sup_{D \in \mathcal{D}} \min_{G \in \mathcal{G}} V(G, D).$$

It is easy to check that Lemma 6 holds even in the non-realizable case. As a result, the generator is minimizing $JSD(p_{data} \| p_G)$ whose value is finite. Clearly the quantities above are finite. Thus there exists a real number v, the value of the game, such that:

$$\begin{aligned} \forall D \in \mathcal{D} &: V(G^*, D) \leq v = V(G^*, D^*) \quad (A) \\ \forall G \in \mathcal{G} &: V(G, D^*) \geq v = V(G^*, D^*) \quad (B) \end{aligned}$$

for G^* the minimizer of $JSD(p_{data} \| p_G)$ and a $D^* \in [0, 1]^{|N|}$. Now applying Lemma 4, we have that

$$\tilde{D} = \text{Best-Response}(G^*) = \frac{p_{data}}{p_{data} + p_{G^*}}.$$

Additionally, by the optimality of the response and the consequence (A) of Minimax Theorem it holds that $V(G^*, \tilde{D}) = v$. Finally, since $V(G^*, \cdot)$ is strongly concave, all other discriminators receive value less than v and are not optimal. Thus

$$D^* = \tilde{D} = \frac{p_{data}}{p_{data} + p_{G^*}}$$

□

D.1.2 f-GAN

The utility of the zero-sum game $V(G, D)$ for the distribution p_{data} over the discrete set N is

$$V(G, D) = \sum_{x \in N} p_{data}(x) D(x) - \sum_{x \in N} p_G(x) f^*(D(x))$$

We will assume that f is a strictly convex function with $f(1) = 0$. On the one hand, it is easy to check that for a fixed discriminator D, the utility function is linear over the p_G operator. On the other hand, for a fixed generator G, the utility function is of the form $aD - bf^*(D)$ which is strictly-concave.

We start our work with the following lemmas

Lemma 9 ([50]). For a fixed generator G the optimal discriminator is

$$D_G(x) = f'\left(\frac{p_{data}(x)}{p_G(x)} \right)$$

Proof. Observe that the optimization problem for each $D(x)$ is decoupled. Thus

$$D_G(x) = \arg\max_D \; p_{data}(x) D - p_G(x) f^*(D)$$

By concavity the unique maximum of the above is given by Fermat criterion

$$D_G(x) = ((f^*)')^{-1}\left(\frac{p_{data}(x)}{p_G(x)} \right) = f'\left(\frac{p_{data}(x)}{p_G(x)} \right)$$

□

[5] We note that D^*, may take the value 1 for some $x \in N$ if the generator G^* does not have full support. Assigning $D(x) = 1$ for some x may lead to infinite utilities in general. We prove however for that the pair (G^*, D^*) this is not the case. We thus consider that pair an equilibrium.

Lemma 10. For a fixed discriminator D, any distribution supported only on

$$S_{G_D} = \{x \in N : \forall x' \in N \quad f^*(D(x)) \geq f^*(D(x'))\}$$

is an optimal generator when it is allowed to choose any distribution over N.

Proof. Observe that for a fixed discriminator, the optimal generator optimizes

$$\sum_{x \in N} p_G(x) f^*(D(x))$$

since the other term is independent of the generator. Let us define the following

$$F_{max} = \max_{x \in N} f^*(D(x))$$

Then we have that

$$\sum_{x \in N} p_G(x) f^*(D(x)) \leq F_{max}$$

with the equality being true only for distributions supported only on S_{G_D}. □

Lemma 11 ([50])**.** The min-max generator is the following distribution

$$G^* = \arg\min_{G \in \mathcal{G}} D_f(p_{data} \| p_G).$$

Proof. We can substitute in $V(G; D)$ the optimal discriminator from Lemma 9. Thus we get

$$V(G; D_G) = \sum_{x \in N} p_{data}(x) f'\left(\frac{p_{data}(x)}{p_G(x)}\right) - \sum_{x \in N} p_G(x) f^*\left(f'\left(\frac{p_{data}(x)}{p_G(x)}\right)\right)$$

We will first prove that:

$$V(G; D_G) = D_f(p_{data} \| p_G)$$

Let's recall firstly the definition of f-divergence:

$$D_f(p_{data} \| p_G) = \sum_{x \in N} p_G(x) f\left(\frac{p_{data}(x)}{p_G(x)}\right)$$

Since f is convex and lower semi-continuous, Frenchel convex duality guarantees that we can write f in terms of its conjugate dual as $f(u) = \sup_{v \in \mathbb{R}} \{uv - f^*(v)\}$. Equivalently we get:

$$\begin{aligned}
D_f(p_{data} \| p_G) &= \sum_{x \in N} p_G(x) \sup_{v \in \mathbb{R}} \left\{\frac{p_{data}(x)}{p_G(x)} v - f^*(v)\right\} \\
&= \sum_{x \in N} \sup_{v \in \mathbb{R}} \{p_{data}(x) v - f^*(v) p_G(x)\} \\
&= \sum_{x \in N} p_{data}(x) f'\left(\frac{p_{data}(x)}{p_G(x)}\right) - \sum_{x \in N} p_G(x) f^*\left(f'\left(\frac{p_{data}(x)}{p_G(x)}\right)\right)
\end{aligned}$$

The last line follows arguments similar to Lemma 9 applied for each term. By minimizing $V(G, D_G)$, the result follows trivially. □

Lemma 12. The max-min discriminator is

$$\forall x \in N : D^*(x) = f'(1)$$

when the generator is allowed choose any distribution over N.

Proof. We want to substitute in $V(G, D)$ the optimal generator from Lemma 5. Observe that for all $x \in S_{G_D}$, we may not have all $D(x)$ to be equal. Only the values of f are guaranteed to be equal, $f(D(x)) = F_{max}$. However, if there are two distinct D values then we can always pick the higher one and improve utility. Thus we can focus on discriminators that are constant over S_{G_D}. Let $D_{F_{max}}$ be the corresponding value

$$V(G_D, D) = D_{F_{max}} \sum_{x \in S_i} p_{data}(x) - f(D_{F_{max}}) \sum_{x \in S_i} p_G(x)$$
$$+ \sum_{x \notin S_{G_D}} p_{data}(x) D(x) - \sum_{x \notin S_{G_D}} \underbrace{p_G(x)}_{0} f(D(x))$$

Let us define $D_{small} = \{D(x) | x \notin S_{G_D}\}$. Observe that if $|D_{small}| > 1$ then setting $D(x) = \max(D_{small})$ improves utility. Thus for the optimal discriminator we have $|D_{small}| = 1$. Let us call $D_{F_{min}}$ the unique element of D_{small}. So for an optimal discriminator we would have a single value $D_{F_{min}}$ with $f(D_{F_{min}}) < f(D_{F_{max}})$. As a result

$$x \notin S_{G_D} \implies D(x) = D_{F_{min}}$$
$$x \in S_{G_D} \implies D(x) = D_{F_{max}}$$

We now have two cases. For any combination with $D_{F_{min}} > D_{F_{max}}$, the constant discriminator $D(x) = D_{F_{min}}$ has higher utility. Symmetrically, for any combination with $D_{F_{max}} > D_{F_{min}}$, the constant discriminator $D(x) = D_{F_{max}}$ has higher utility. Thus the optimal discriminator is constant. Plugging in the constant discriminator $D_{const}(x) = D$ we get

$$V(G_{D_{const}}, D_{const}) = D + f(D)$$

The optimal value for D follwoing the approach of Lemma 9 is $f'(1)$ and as a result

$$D(x) = f'(1)$$

□

Lemma 13 (Non-realizable case). Assume that $f \in C^1$ is strictly convex and $\lim_{x \to 0^+} x f(\frac{1}{x})$ exists and is finite[6]. If the choice of generator G is restricted in $\mathcal{G}$, a convex compact subset of the $|N|$ dimensional simplex, such that $p_{data} \notin \mathcal{G}$ then

$$(G, D) = \left(\arg\min_{G \in \mathcal{G}} D_f(p_{data} || p_G), f'\left(\frac{p_{data}}{p_G}\right)\right)$$

Proof. We cannot readily apply von Neumann's minimax theorem since the $V(G, D)$ since $\mathcal{D} = \mathbb{R}^{|N|}$ is not compact for the discriminator. We can still apply Fan's Minimax Theorem

$$\min_{G \in \mathcal{G}} \sup_{D \in \mathcal{D}} V(G, D) = \sup_{D \in \mathcal{D}} \min_{G \in \mathcal{G}} V(G, D).$$

It is easy to check that Lemma 12 holds even in the non-realizable case. As a result, the generator is minimizing $D_f(p_{data} || p_G)$ whose value is finite under the assumptions we made on f. Clearly the quantities above are finite. Thus there exists a real number v, the value of the game, such that:

$$\forall D \in \mathcal{D} : V(G, D) \leq v = V(G, D) \quad (A)$$
$$\forall G \in \mathcal{G} : V(G, D) \geq v = V(G, D) \quad (B)$$

for G the minimizer of $D_f(p_{data} || p_G)$ and a $D \in \mathbb{R}^{|N|}$. Now applying Lemma 9 we have that

$$\tilde{D} = \text{Best-Response}(G) = f'\left(\frac{p_{data}(x)}{p_G(x)}\right).$$

Additionally, by the optimality of the response and the consequence (A) of Minimax Theorem it holds that $V(G, \tilde{D}) = v$. Finally, assuming that f is strictly convex we get that $V(G, \cdot)$ is strictly concave, Best-Response(G) is unique and thus

$$D = \tilde{D} = f'\left(\frac{p_{data}(x)}{p_G(x)}\right)$$

□

[6] This assumption guarantees that the D_f is always finite even if the distribution chosen by the generator is not fully supported on N. This in turn guarantees that D is also finite resulting in a meaningful equilibrium. Unbounded divergences like KL are known to be problematic for GANs even in practice [4].

D.1.3 WGAN

The utility of the zero-sum game $V(G;D)$ for the distribution p_{data} over the discrete metric space $(\mathcal{N};dist)$

$$V(G;D) = \mathbb{E}_{X\sim p_{data}}[D(X)] - \mathbb{E}_{X\sim p_G}[D(X)]$$
$$= \sum_{x\in\mathcal{N}} (p_{data}(x) - p_G(x))D(x) \text{ where } \|D\|_{Lip} \le 1$$

On the one hand, it is easy to check that for a fixed discriminator D, the utility function is linear over the p_G operator. On the other hand, for a fixed generator G, the utility function is linear over D.

We start our work with the following lemmas

Lemma 14 ([4])**.** For a fixed generator G the optimal discriminator is a solution of the following linear program

$$\text{maximize over } D(\cdot) \quad \sum_{x\in\mathcal{N}} (p_{data}(x) - p_G(x))D(x)$$
$$\text{subject to} \quad |D(x) - D(x')| \le dist(x;x'); \forall x;x' \in \mathcal{N}$$

where the optimal value of the LP is the Earth mover's distance between p_{data} and p_G.

Proof. Indeed, by definition any solution of the above LP is an optimal discriminator over a fixed generator G. To complete the proof of the statement, we recall that Earth Mover's distance of $(p_{data};p_G)$ is equal to

$$\min_{\gamma\in Coupling(p_{data};p_G)} \mathbb{E}_{(X;X')\sim\gamma}[dist(X;X')]:$$

Now if we consider the dual formulation of the Wasserstein distance, then the Kantorovich duality [21, 64] implies that the above linear program consists exactly the dual linear program which computes the Earth Mover's distance. □

Lemma 15. For a fixed discriminator D, any distribution supported only on

$$S_{G_D} = \{x \in \mathcal{N} : \forall x' \in \mathcal{N} \quad D(x) \ge D(x')\}$$

is an optimal generator when it is allowed to choose any distribution over $\mathcal{N}$.

Proof. Observe that for a fixed discriminator, the optimal generator optimizes

$$\sum_{x\in\mathcal{N}} p_G(x)D(x)$$

since the other term is independent of the generator. Let us define the following

$$D_{max} = \max_{x\in\mathcal{N}} D(x)$$

Then we have that
$$\sum_{x\in\mathcal{N}} p_G(x)D(x) \le D_{max}$$

with the equality being true only for distributions supported only on S_{G_D}. □

Lemma 16 ([4])**.** The min-max generator is the following distribution

$$G^* = \arg\min_{G\in\mathcal{G}} EMD(p_{data};p_G):$$

Proof. We can substitute in $V(G;D)$ the optimal discriminator from Lemma 14. Thus we get

$$V(G;D_G^*) = EMD(p_{data};p_G)$$

By minimizing $V(G;D_G^*)$, the result follows trivially. □

Lemma 17. The max-min discriminator is

$$\forall x \in \mathcal{N} : D^*(x) = c, \text{ Constant function}$$

when the generator is allowed choose any distribution over $\mathcal{N}$.

Proof. We can substitute in $V(G, D)$ the optimal generator from Lemma 5

$$V(G_D, D) = D_{max} \sum_{x \in S_{G_D}} p_{data}(x) - D_{max} \sum_{x \in S_{G_D}} p_G(x) + \sum_{x \notin S_{G_D}} p_{data}(x) D(x) - \sum_{x \notin S_{G_D}} \underbrace{p_G(x)}_{0} D(x)$$

Observe that for $x \notin S_{G_D}$, if D takes more than two values then setting D equal to the highest of the them for all $x \notin S_{G_D}$ improves utility. So for an optimal discriminator we would have a single value $D_{max} > D_{min}$. In the end we have that

$$x \notin S_{G_D} \implies D(x) = D_{min}$$
$$x \in S_{G_D} \implies D(x) = D_{max}$$

Observe that for any combination of D_{max} and D_{min} with $D_{max} > D_{min}$, the constant discriminator D_{max} has higher utility. Therefore we can focus our attention on the constant discriminator $D_{const}(x) = D$, where the optimal value is exactly zero.

$$V(G_{D_{const}}, D_{const}) = 0$$

Finally, it is easy to check that the choice of constant discriminator satisfies trivially the Lipschitz constraints, i.e. $|D_{const}(x) - D_{const}(x')| = 0 \leq dist(x, x')$ for any metric function $dist$. □

D.2 GANs and Hidden Constrained Optimization

In the following section, we will generalize the results of Section 3.2 and Section 3.3 for the case of vanilla GAN of [26] whose objective is linear-strong-concave where the maximization part is constrained in the distributional simplex. More precisely,

$$\min_{\substack{p_G(x) \geq 0; \\ \sum_{x \in \mathcal{N}} p_G(x) = 1}} \max_{D \in (0,1)^{|\mathcal{N}|}} V(G, D) = \sum_{x \in \mathcal{N}} p_{data}(x) \log(D(x)) + \sum_{x \in \mathcal{N}} p_G(x) \log(1 - D(x))$$

At a first glance, by rewriting the equivalent Langrangian formulation of the aforementioned constrained min-max problem we can see that strong-concavity property does not hold any more. However our following theorem shows that by exploiting further the structure of [26]'s architecture a convergence result is possible.

Theorem 16. Let $V(\text{Gen}_\theta, \text{Disc}_\phi)$ be Goodfellow GAN as described in Section 4, where G, D use sigmoid activations. Then for a fully mixed distribution p_{data}, $(F(t) = \text{Gen}_{\theta(t)}, G(t) = \text{Disc}_{\phi(t)})$ converges to $(p_{data}, \frac{1}{2} 1_{|\mathcal{N}|})$ as $t \to \infty$ under the dynamics of Equation (1).

Proof. Let us write down our original objective

$$\min_{\substack{p_G(x) \geq 0; \\ \sum_{x \in \mathcal{N}} p_G(x) = 1}} \max_{D \in (0,1)^{|\mathcal{N}|}} V(G, D) = \sum_{x \in \mathcal{N}} p_{data}(x) \log(D(x)) + \sum_{x \in \mathcal{N}} p_G(x) \log(1 - D(x))$$

In order to remove the constraints from the objective above, we plan to make use of a Lagrange multiplier. We remind the reader that since both the discriminator and the generator use the sigmoid activations, we only have to capture the $\sum_{x \in \mathcal{N}} p_G(x) = 1$ constraint. Thus, our equivalent Langragian is:

$$\min_{\theta \in \mathbb{R}^{|\mathcal{N}|}} \max_{\phi \in \mathbb{R}^{|\mathcal{N}|};\ \lambda \in \mathbb{R}} L(F, G, \lambda) = p_{data}^\top \log(G(\phi)) + F(\theta)^\top \log(1 - G(\phi)) + \lambda (F(\theta)^\top 1_{|\mathcal{N}|} - 1)$$

where

$$\begin{aligned} F(\theta) &= f_1(\theta_1) \times f_2(\theta_2) \times \cdots \times f_{|N|}(\theta_{|N|}) \\ G(\phi) &= g_1(\phi_1) \times g_2(\phi_2) \times \cdots \times g_{|N|}(\phi_{|N|}) \end{aligned}$$

and f_i and g_j are sigmoid functions and θ_i and ϕ_j are their one dimensional inputs. Let's write again the equivalent dynamics of Equation (3) for the sigmoid activations and the Langrage multiplier. Applying the same steps with Theorem 4 for sigmoids:

$$\begin{cases} \dot{f}_i = -f_i^2(1-f_i)^2 \frac{\partial L}{\partial f_i}(F,G) & \forall i \in [|N|] \\ \dot{g}_j = g_j^2(1-g_j)^2 \frac{\partial L}{\partial g_j}(F,G) & \forall j \in [|N|] \\ \dot{\lambda} = \sum_{i=0}^{|N|} f_i - 1 \end{cases}$$

Since all initializations are safe in this game, our "generalized" Lyapunov function:

$$H(F,G,\lambda) = \sum_{i=0}^{|N|} \int_{p_{data}(x_i)}^{f_i} \frac{z - p_{data}(x_i)}{z^2(1-z)^2} dz + \sum_{j=0}^{|N|} \int_{1/2}^{g_j} \frac{z - 1/2}{z^2(1-z)^2} dz + \frac{(\lambda - \lambda^*)^2}{2}$$

where λ^* is the Langrange multiplier at the equilibrium of the non-hidden game and x_i is the i-th element of N. Applying the same steps as in Lemma 3 we get that GDA approaches the largest invariant set E of points (F,G,λ) that have the following properties

$$L(p_{data}, G, \lambda) = L\left(p_{data}, \frac{1}{2}\mathbf{1}_{|N|}, \lambda^*\right)$$

$$L\left(F, \frac{1}{2}\mathbf{1}_{|N|}, \lambda^*\right) = L\left(p_{data}, \frac{1}{2}\mathbf{1}_{|N|}, \lambda^*\right)$$

For the first equality, we have that the value of λ does not affect L when the generator respects the sum to one constraint. Thus

$$L(p_{data}, G, \lambda) = L(p_{data}, G, \lambda^*)$$

Then we can observe that $L(p_{data}, G, \lambda^*)$ is strictly concave in G and given that $\frac{1}{2}\mathbf{1}_{|N|}$ is its unique minimum we have that

$$L(p_{data}, G, \lambda^*) = L\left(p_{data}, \frac{1}{2}\mathbf{1}_{|N|}, \lambda^*\right) \implies G = \frac{1}{2}\mathbf{1}_{|N|}$$

Given that E is an invariant set and G is constant in E, we have that $\dot{G} = 0$. In other words,

$$0 = \frac{1}{2^2}\left(1 - \frac{1}{2}\right)^2 \frac{\partial L}{\partial g_j}\left(F, \frac{1}{2}\mathbf{1}_{|N|}, \lambda\right) \quad \forall j \in [|N|]$$

As a consequence we have that

$$\frac{\partial L}{\partial g_j}\left(F, \frac{1}{2}\mathbf{1}_{|N|}, \lambda\right) = 0 \implies f_j = p_{data}(x_j) \quad \forall j \in [|N|]$$

Once again, given that E is an invariant set and F is constant in E, we have that $\dot{F} = 0$

$$0 = p_{data}(x_i)^2 (1 - p_{data}(x_i))^2 \frac{\partial L}{\partial f_i}\left(p_{data}, \frac{1}{2}\mathbf{1}_{|N|}, \lambda\right) \quad \forall i \in [|N|]$$

This leads to

$$\frac{\partial L}{\partial f_i}\left(p_{data}, \frac{1}{2}\mathbf{1}_{|N|}, \lambda\right) = 0 \implies \lambda = \log \frac{1}{2} \quad \forall i \in [|N|]$$

Observe that by the optimality conditions of the non-hidden game, λ^* needs to satisfy the same equation and thus $\lambda = \lambda^*$. Clearly we have that

$$(F,G,\lambda) \in E \implies (F,G,\lambda) = \left(p_{data}, \frac{1}{2}\mathbf{1}_{|N|}, \lambda^*\right)$$

Thus the dynamics converge to the unique equilibrium of the hidden game. □

D.3 Zero-Sum Games

We close this section with an application of our regularization machinery in hidden bilinear games. Hidden bilinear zero-sum games were introduced by [65] and they are formally defined as:

Definition 11 (Hidden Bilinear Zero-Sum Game)**.** *In a hidden bilinear zero-sum game there are two players, each one equipped with a smooth function* $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^N$ *and* $\boldsymbol{G} : \mathbb{R}^m \to \mathbb{R}^M$ *and a payoff matrix* $U_{N \times M}$ *such that each player inputs its own decision vector* $\theta \in \mathbb{R}^n$ *and* $\phi \in \mathbb{R}^m$ *and is trying to maximize or minimize* $r(\theta, \phi) = \boldsymbol{F}(\theta)^\top U \boldsymbol{G}(\phi)$ *respectively.*

For the special case of hidden bilinear games, [65] proved that if the dimension of the game is greater or equal than two like (e.g. akin to Rock-Paper-Scissors) then GDA dynamics tend to "cycle" through their parameter space with an even more complex behavior than a typical periodic trajectory. Specifically, the system is formally analogous to Poincaré recurrent systems (e.g. many body problem in physics). In contrast, leveraging Theorem 7, we know that by adding a small regularization term we can "break" the cycling behavior and converge to an approximate Nash Equilibrium. We close this section by presenting a comparison between the optimization portraits of GDA dynamics with the absence or not of a regularization for the archetypical game of Rock-Paper-Scissors:

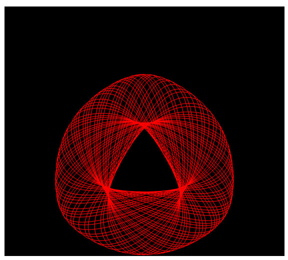

Figure 6: Trajectories of a single player using gradient-descent-ascent dynamics for a hidden bilinear game $L(\mathrm{F}(\theta), \mathrm{G}(\phi)) = \mathrm{F}^\top(\theta) A \mathrm{G}(\phi)$ where A is the classical Rock-Paper-Scissors table and F, G have the sigmoid activations. The two left figures present the Poincaré recurrence for different initializations of the dynamics, a behavior consistent with the Lyapunov stability of Theorem 3. On the other hand, the two figures on the right illustrate convergent to the mixed Nash equilibrium executions which exploit the regularization tools as described in Section 3.3. The regularization terms added are centered at the mixed equilbrium of the game, leading to convergence to the unmodified equilibrium of the Rock-Paper-Scissors game.

Remark 5. *Closing this appendix, it would be usefule to clarify some details between this work and [65]. While we use tools regarding reparametrization and safety from their work, the rest of our analysis and the technical ideas behind them are qualitatively different. [65] uses the Poincaré recurrence theorem to argue that hidden bi-linear games exhibit recurrent behavior even under safety. In contrast we show that these games are* **merely edge cases** *and that for strictly convex concave HCC convergence to the underlying Von Neumann solution is guaranteed for safe initial conditions. To the best of our knowledge, our work is the first one to provide* **sufficient conditions for non-local convergence** *to a game theoretically meaningful solution for a wide family of* **non-convex non-concave min-max** *problems. In addition, [65] require* F, G *to be invertible operators in order to prove recurrence in the input space. In contrast, in our work we do not rely on invertibility to transfer convergence from the output space dynamics of Equation* (3) *to the input dynamics of Equation* (1).