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

Zixuan Pan
zxp@umich.edu

Abstract

Many of the recent advances in machine learning have been inspired by min-max games, however many of this methods lacks a solid math background telling about their dynamics. Whether these methods could always converge, and which point they will converge to remains much unexplained. This paper propose a formulation for a kind of games called hidden convex concave games, which is related to more general min-max games. The paper proposed two main results:if the game is eequipped with a strictly convex concave objective function, then the convergence to a von Neumann solution can be guaranteed by designing a Lyapunov function. Meanwhile, stronger regularization can make convergence faster, but could also make the convergence point shift. Also some implementation of this paper's result on real datasets is provided in this review.

1. Intoduction

Many of the recent advances in machine learning have been inspired by min-max games, such as generative adversarial networks (GAN) in image generation, and actor-critic algorithm in reinforcement learning [1, 4, 8]. These problems are often solved by applying gradient descent ascent (GDA) to some neural network based models. As a main feature of neural networks, the optimization problems could always be non-convex non-concave, but they are bounded by a convex-concave loss function. This kind of min-max game is defined as Hidden Convex-Concave (HCC) games [3]. A formal mathematical definition of it could be: given $\boldsymbol{F} : \mathbb{R}^N \to X \subset \mathbb{R}^n$ and $\boldsymbol{G} : \mathbb{R}^M \to Y \subset \mathbb{R}^m$ and a continuous convex-concave function $L : X \times Y \to \mathbb{R}$, such that the min-max game is $\min_{\boldsymbol{\theta}\in\mathbb{R}^N} \max_{\boldsymbol{\phi}\in\mathbb{R}^M} L(\boldsymbol{F}(\boldsymbol{\theta}), \boldsymbol{G}(\boldsymbol{\phi}))$.

Studying the dynamics of GDA algorithms in HCC games is critical, as there's a bunch of underlying issues that are related to unstable training of HCC games. The vanilla GAN requires heavily parameter tuning to make the training process stable [1]. Meanwhile, the GDA algorithm may often fall into some meaningless saddle points, which generates undesired solutions. WGAN solves part of the problems by making modifications to the loss function, but it still needs some tricks like gradient clipping or adding extra regularization terms to stabilize the training process [5], which still needs proper parameter tuning. Thus, it's very useful to study the dynamics of HCC games and find how to make GDA converge to a meaningful minima. This could shed light on the advances of such min-max game based models by reducing the parameter tuning work with explainability.

This paper focused on finding appropriate initialization to guarantee stability of the dynamic system and the relation between convergence and regularization.

2. Preliminary

2.1. Hidden Convex Concave Games

We will start with giving a formal definition of HCC games.

Definition 1 (HCC games) *$L : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ 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.*

As the property of min-max games, we assume two players are involved in the game (solving the optimization problem). The goal of one player is to minimize the objective function L, while the goal of the other player is to maximize L. We assume 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 $g_j : \mathbb{R}^{m_j} \to \mathbb{R}$. In order to study the dynamic, we will assume that f_i, g_j, L are all C^2 functions, which means twice differentiable. The inputs $\boldsymbol{\theta}_i \in \mathbb{R}^{n_i}$ and $\boldsymbol{\phi}_j \in \mathbb{R}^{m_j}$ are grouped in two vectors

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \cdots & \boldsymbol{\theta}_n \end{bmatrix}^\top \quad \boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\boldsymbol{\theta}_1) & f_2(\boldsymbol{\theta}_2) & \cdots & f_N(\boldsymbol{\theta}_n) \end{bmatrix}^\top$$

$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \cdots & \boldsymbol{\phi}_m \end{bmatrix}^\top \quad \boldsymbol{G}(\boldsymbol{\theta}) = \begin{bmatrix} g_1(\boldsymbol{\phi}_1) & g_2(\boldsymbol{\phi}_2) & \cdots g_M(\boldsymbol{\phi}_m) \end{bmatrix}^\top$$

In the view of neural network, the ϕ and θ can be viewed as some high dimensional inputs of neural networks, while F and G are two different networks.

Therefore, the hidden convex concave game is:

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

where $N = \Sigma_{i=1}^n n_i$ and $N = \Sigma_{i=1}^m m_i$

2.2. Dynamics of HCC games

Definition 2 (Von Neumann Solutions) *Given a convex concave function L, we denote the stationary points of L as (global) Nash equilibria of the min-max game. The set of all equilibria is defined as von Neumann solutions of L, denote as Solution(L).*

When the game is defined on a convex compact set, existence of at least a solution is guaranteed by von Neumann's minimax theorem [2]. By choosing appropriate f_i and g_i, the convex compact setting can be obtained, and thus in the following section we assume there's Solution(L) is not empty. Then the gradient descent ascent dynamic of HCC games can be defined as follow:

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

2.3. Reparametrization

The stability of HCC games is guaranteed under appropriate initialization, which will be discussed later. Here we'd like to make some definitions of the initialization and solutions of dynamic system first.

Lemma 3 (Unique Solution of Dynamic System) *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 $\boldsymbol{x}(t)$ denote the unique solution of the dynamical system Σ_1. Then the unique solution for dynamical system Σ_2 is $\boldsymbol{z}(t) = \boldsymbol{x}\left(\int_0^t h(s)\mathrm{d}s\right)$*

$$\left\{ \begin{array}{ccc} \dot{\boldsymbol{x}} & = & \nabla k(\boldsymbol{x}) \\ \boldsymbol{x}(0) & = & \boldsymbol{x}_{init} \end{array} \right\} : \Sigma_1 \quad \left\{ \begin{array}{ccc} \dot{\boldsymbol{z}} & = & h(t)\nabla k(\boldsymbol{z}) \\ \boldsymbol{z}(0) & = & \boldsymbol{x}_{init} \end{array} \right\} : \Sigma_2$$

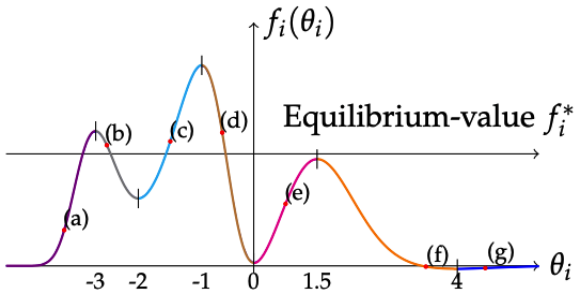

Figure 1. Relation between initialization and GDA trajectory.

See Appendix A for proof.

By choosing $h(t) = -\partial L(\boldsymbol{F}(t), \boldsymbol{G}(t))/\partial f_i$ and $h(t) = \partial L(\boldsymbol{F}(t), \boldsymbol{G}(t))/\partial g_j$ respectively, we can connect the dynamics of each $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_j$ under Equation (1) to gradient ascent on f_i and g_j. Applying Lemma 1, we get that trajectories of $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_j$ under Equation (1) are restricted to be subsets of the corresponding gradient ascent trajectories with the same initializations. In fact, Lemma 1 guarantees the monotomicity of GDA trajectories. As shown 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.

Definition 4 (Image of Dynamic System) *For each initialization $\boldsymbol{x}(0)$ of $\Sigma_1, \mathrm{Im}_k(\boldsymbol{x}(0))$ is the image of $k \circ \boldsymbol{x} : \mathbb{R} \to \mathbb{R}$.*

Applying Definition 2 in Figure 1, $\mathrm{Im}_{f_i}(\theta_i(0)) = (f_i(-2), f_i(-1))$ 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 $\boldsymbol{\theta}_i$ are vectors, Lemma 1 implies that for a given $\boldsymbol{\theta}_i(0)$, $f_i(\boldsymbol{\theta}_i(t))$ uniquely identifies $\boldsymbol{\theta}_i(t)$. As a result we get that a new dynamical system involving only f_i, g_j and initializations.

Theorem 5 (GDA without Intermediate States) *For each initialization $(\boldsymbol{\theta}(0), \boldsymbol{\phi}(0))$ of Equation (1), there are C^1 functions $X_{\boldsymbol{\theta}_i(0)}, X_{\boldsymbol{\phi}_j(0)}$ such that $\boldsymbol{\theta}_i(t) = X_{\boldsymbol{\theta}_i(0)}(f_i(t))$ and $\boldsymbol{\phi}_j(t) = X_{\boldsymbol{\phi}_j(0)}(g_j(t))$. If $(\boldsymbol{\theta}(t), \boldsymbol{\phi}(t))$ satisfy Equation (1) then $f_i(t) = f_i(\boldsymbol{\theta}_i(t))$ and $g_j(t) = g_j(\boldsymbol{\phi}_j(t))$ satisfy*

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

The proof is given in Appendix B.

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 6 (Safe Initialition) *The initialization $(\boldsymbol{\theta}(0), \boldsymbol{\phi}(0))$ is called safe for a $(\boldsymbol{p}, \boldsymbol{q}) \in$ Solution (L) if $\boldsymbol{\phi}_i(0)$ and $\boldsymbol{\theta}_j(0)$ are not stationary points of f_i and g_j respectively and $p_i \in \mathrm{Im}_{f_i}(\boldsymbol{\theta}_i(0))$ and $q_j \in \mathrm{Im}_{g_j}(\boldsymbol{\phi}_j(0))$.*

2.4. Stability of Dynamic Systems

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

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) \tag{$\star$}$$

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

Definition 7 (Stability properties [7]) *The fixed point $\boldsymbol{x} = \mathbf{0}$ of Equation ($\star$) is*
stable if, for each $\epsilon > 0$, there is a $\delta = \delta(\epsilon) > 0$ such that

$$\|\boldsymbol{x}(0)\| < \delta \Longrightarrow \|\boldsymbol{x}(t)\| < \epsilon \quad \forall t \geq 0$$

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

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

Lyapunov Function playes an important role in determining the stability of fixed points of a dynamic system.

Theorem 8 (Lyapunov Theorem [7]) *Let $\boldsymbol{x} = \mathbf{0}$ be a fixed point point for Equation ($\star$) and $D \subset \mathbb{R}^n$ be a domain containing $\boldsymbol{x} = \mathbf{0}$. Let $V : D \to \mathbb{R}$ be a continuously differentiable function such that*

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

then $\boldsymbol{x} = \mathbf{0}$ is stable. Moreover if

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

then $\boldsymbol{x} = \mathbf{0}$ is asymptotically stable.

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

3. Lyapunov Function for GDA of HCC Games

3.1. General Case

The section will be about designing a Lyapunov Function for Equation (2).

Lemma 9 *If L is convex concave and $(\phi(0), \boldsymbol{\theta}(0))$ is a safe for $(\boldsymbol{p}, \boldsymbol{q}) \in$ Solution (L), then the following quantity is non-increasing under the dynamics of Equation (3):*

$$H(\boldsymbol{F}, \boldsymbol{G}) = \sum_{i=1}^{N} \int_{p_i}^{f_i} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \,\mathrm{d}z + \sum_{j=1}^{M} \int_{\boldsymbol{q}_j}^{g_j} \frac{z - q_j}{\left\|\nabla g_j\left(X_{\boldsymbol{\phi}_j(0)}(z)\right)\right\|^2} \,\mathrm{d}z$$

See Appendix C for proof.

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

The last inequality holds since $(\boldsymbol{p},\boldsymbol{q}) \in \text{Solution}(L)$.
If $(\boldsymbol{p},\boldsymbol{q})$ is a saddle point of L then $L(\boldsymbol{p},\boldsymbol{G}) \leq L(\boldsymbol{p},\boldsymbol{q}) \leq L(\boldsymbol{F},\boldsymbol{q})$

Theorem 10 *If L is convex concave and $(\boldsymbol{\phi}(0),\boldsymbol{\theta}(0))$ is a safe for $(\boldsymbol{p},\boldsymbol{q}) \in \text{Solution}(L)$, then $(\boldsymbol{p},\boldsymbol{q})$ is stable for Equation (2).*

See Appendix D for proof.

Theorem 11 (Safety properties of Sigmoid Functions) *If f_i and g_j are sigmoid functions and L is convex concave and there is $a(\boldsymbol{\phi}(0),\boldsymbol{\theta}(0))$ that is safe for $(\boldsymbol{p},\boldsymbol{q}) \in \text{Solution}(L)$, then $\left(\boldsymbol{F}^{-1}(\boldsymbol{p}),\boldsymbol{G}^{-1}(\boldsymbol{q})\right)$ is stable for Equation (1).*

See Appendix 11 for Proof.
With these theorems, the stability of such a dynamic system is guaranteed with a pair (p,q) in the solution space.

Theorem 12 (Bound of HCC games) *Let $(\boldsymbol{p},\boldsymbol{q}) \in \text{Solution}(L)$. Let R_{f_i} and R_{g_j} be the set of regular values of f_i and g_j respectively. Assume that there is a $\xi > 0$ such that $[p_i - \xi, p_i + \xi] \subseteq R_{f_i}$ and $[q_j - \xi, q_j + \xi] \subseteq R_{g_j}$. Define*

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

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

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

See Appendix F for the proof.
This theorem provides that the dynamic system over time is bounded. However, it does not guarantee convergence to a certain point. Thus, we will further discuss L as a strictly convex concave function.

3.2. Hidden Strictly Convex Concave Games

Here we continue to study properties of hidden strictly convex concave games with L as a strictly convex concave function.

Lemma 13 (Locally Asymptotically Stability of Strictly Convex Concave Functions) *Let L be strictly convex concave and $Z \subset \text{Solution}(L)$ is the non empty set of equilbria of L for which $(\boldsymbol{\theta}(0),\boldsymbol{\phi}(0))$ is safe. Then Z is locally asymptotically stable for Equation (2).*

See Appendix G for proof.

Theorem 14 (Convergence of Hidden Strictly Convex Concave Games) *Let L be strictly convex concave and $\text{Z} \subset \text{Solution}(L)$ is the non empty set of equilbria of L for which $(\boldsymbol{\theta}(0),\boldsymbol{\phi}(0))$ is safe. Under the dynamics of Equation (1)$(\boldsymbol{F}(\boldsymbol{\theta}(t)),\boldsymbol{G}(\boldsymbol{\theta}(t)))$ 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 15 *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.*

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

3.3. Convergence via Regularization

Regularization has become a widely used technique in neural networks and also in GANs [1, 4]. It could also have impact on HCC games. Assume a utility $L(\boldsymbol{x}, \boldsymbol{y})$ that is convex concave but not strictly. Here we will propose a modified utility L' that is strictly convex strictly concave. Specifically we will choose

$$L'(\boldsymbol{x}, \boldsymbol{y}) = L(\boldsymbol{x}, \boldsymbol{y}) + \frac{\lambda}{2}\|\boldsymbol{x}\|^2 - \frac{\lambda}{2}\|\boldsymbol{y}\|^2$$

The choice of the parameter λ 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 λ the equilibria of L are not significantly perturbed.

Theorem 16 (Equilibria of Invertible Hessian) *If L is a convex concave function with invertible Hessians at all its equilibria, then for each* $\epsilon > 0$ *there is a* $\lambda > 0$ *such that* L' *has equilibria that are* ϵ*-close to the ones of L.*

See Appendix I for proof.

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

Theorem 17 (Bounded by Regularization) *Let* $(\boldsymbol{\theta}(0), \boldsymbol{\phi}(0))$ *be a safe initialization for the unique equilibrium of* $L'(\boldsymbol{p}, \boldsymbol{q})$*. If*

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

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

See Appendix J for proof.

This theorem guarantees that the dynamic is bounded by regularization

4. Applications

4.1. Hidden strictly convex-concave games

A famous example of HCC game is GAN. The conclusion of the paper can be applied to all variants of GANs [1, 4, 8]. In the vanilla GAN architecture, as it is commonly referred, our goal is to obtain a generator, such that it can generate a distribution p_{G} that is close to an input data distribution $p_{\text{data.}}$. To find such a generator function, we can use a discriminator D that judge if an input is real or fake. For the case of a discrete p_{data} over a set $\mathcal{N}$, the minimax problem of GAN is the following:

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

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 $p_{\text{G}}(x)$. On the other hand, for a fixed generator $G^*, V(G^*, D)$ is strongly-concave. We can implement the inequality constraints on both the generator probabilities and discriminator using sigmoid activations. For the equality constraint $\sum_{x\in\mathcal{N}} p_{\text{G}}(x) = 1$ we can introduce a Langrange multiplier. Theretically, as given in the paper, when having effectively removed the constraints, we

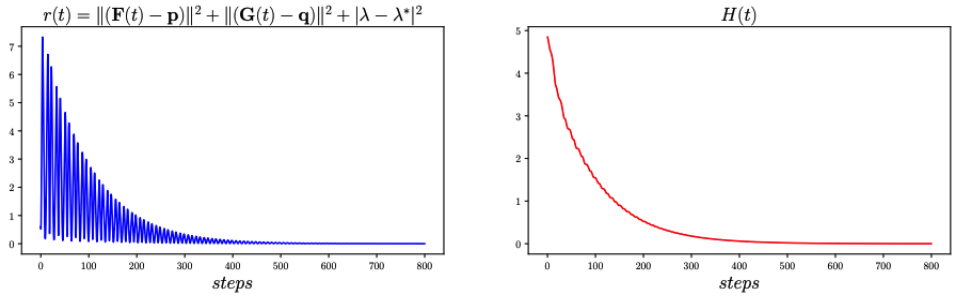

Figure 2. Theoretical l2 distance and Lyapunov function to the equilibria.

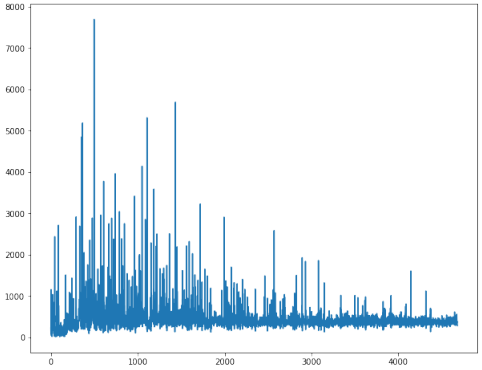

Figure 3. l2 distance on MNIST.

can see in Figure 2, the dynamics of Equation (1) converge to the unique equilibrium of the game. Meanwhile, the dynamic of GAN on MNIST dataset is implemented[1]. An approximation is done to obtain $r(t)$, as p_{data} is hard to obtain, where $\|G(\phi(t)) - q\|$ term is approximated with $\|D(G(x))\|$. See Figure 3 for the l2 distance on MNIST. The result is similar to the theoretical one that l2 distance does not converge monotonically to 0 (left on Figure 2).

4.2. Hidden Convex Concave Games with Regularization

Convergence with respect to regularization strength for vanilla GAN is given in Figure 4. Generally a larger regularization strength gives a faster convergence speed. However, a faster convergence cannot always guarantee converging to a von Neumann solution. Figure 5,6 shows that the generation with larger regularization could fail, indicating not converging to a meaningful solution.

Another example of GAN is WGAN [1]. One of the contributions of 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.The original WGAN formulation has a Lipschitz constraint in the discriminator function which is done by weight clipping. However, model performance is very sensitive to the weight clipping ratio. Thus, [5] replaced this constraint with a quadratic regularizer. The min-max problem for the case of one-dimensional Gaussian $\mathcal{N}\left(0, \alpha_*^2\right)$ and linear discriminator $D_v(x) = v^\top x$ with x^2 activation is:

$$\begin{aligned}\min_{\alpha\in\mathbb{R}} \max_{v\in\mathbb{R}} V_{\mathrm{WGAN}}\left(G_\alpha, D_v\right) &= \mathbb{E}_{\boldsymbol{X}\sim p_{\text{data}}}[D(\boldsymbol{X})] - \mathbb{E}_{\boldsymbol{X}\sim p_{\mathrm{G}}}[D(\boldsymbol{X})] - v^2/2 \\ &= \mathbb{E}_{x\sim\mathcal{N}(0,\alpha_*^2)^2}[vx] - \mathbb{E}_{x\sim\mathcal{N}(0,\alpha_*^2)^2}[vx] - v^2/2 \\ &= \left(\alpha_*^2 - \alpha^2\right) v - v^2/2\end{aligned}$$

[1] See https://github.com/zxp46/EECS559-Final-Project for the code.

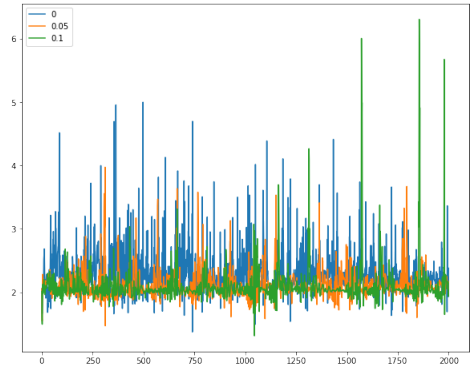

Figure 4. Convergence and regularization strength.

Figure 5. Generation samples for $\lambda = 0$.

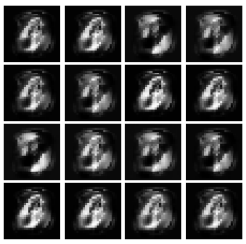

Figure 6. Generation samples for $\lambda = 0.1$.

Observe that V_{WGAN} is not convex-concave but it can posed as a hidden strictly convex-concave game with $\boldsymbol{G}(\alpha) = \left(\alpha_*^2 - \alpha^2\right)$ and $\boldsymbol{F}(v) = v$. When computing expectations analytically without sampling, Theorem 14 still guarantees convergence.

5. Discussion and Conclusion

This paper discusses about the dynamic of HCC games, where the main contribution is proposing that if it is a hidden strictly convex concave game, then the convergence to a meaningful von Neumann solution can be guaranteed by designing a Lyapunov function. Meanwhile, stronger regularization can make convergence faster, but could also make the convergence point shift. For more general cases, as also discussed in [6], the convergence to a meaningful point is not definite. Another limitation of this paper is that the Lyapunov function is hard to implement with code, as the distribution of generated samples and original samples are difficult to obtain. But overall, this paper shed light on formulating a better theoretical background for some zero-sum games.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 1, 6, 7

[2] Hichem Ben-El-Mechaiekh and Robert Dimand. A simpler proof of the von neumann minimax theorem. *American Mathematical Monthly*, 118, 08 2011. 2

[3] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. *arXiv preprint arXiv:2101.05248*, 2021. 1

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1, 6

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 1, 7

[6] Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pages 4337–4348. PMLR, 2021. 8

[7] H.K. Khalil. *Nonlinear Systems*. Pearson Education. Prentice Hall, 2002. 4

[8] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999. 1, 6

A. Proof of Lemma 3

Proof.

$$\boldsymbol{z}(0) = \boldsymbol{x}\left(\int_0^0 h(s)\mathrm{d}s\right) = \boldsymbol{x}(0) = \boldsymbol{x}_{\text{init}}$$

$$\begin{aligned}\dot{\boldsymbol{z}} &= \dot{\boldsymbol{x}}\left(\int_0^t h(s)\mathrm{d}s\right) \times \frac{\mathrm{d}\left[\int_0^t h(s)\mathrm{d}s\right]}{\mathrm{d}t} \\ &= \nabla k\left(\boldsymbol{x}\left(\int_0^t h(s)\mathrm{d}s\right)\right) h(t) = \nabla k(\boldsymbol{z})h(t)\end{aligned}$$

gives a unique solution for Σ_2

B. Proof of Theorem 5

Proof. Let us first study a simpler dynamical system (Σ^*) with unique solution of $\gamma_{\theta_i(0)}(t)$.

$$(\Sigma^*) \equiv \left\{ \begin{array}{ccc} \dot{\boldsymbol{z}} & = & \nabla f_i(\boldsymbol{z}) \\ \boldsymbol{z}(0) & = & \boldsymbol{\theta}_i(0) \end{array} \right\}$$

It is easy to observe that:

$$\dot{f}_i = \nabla f(\boldsymbol{z})\dot{\boldsymbol{z}} = \|\nabla f(\boldsymbol{z})\|^2$$

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

$$X_{\boldsymbol{\theta}_i(0)}\left(f_i\right) = \boldsymbol{\theta}_i(0).$$

On the other hand if $\theta_i(0)$ is not a stationary point of f_i, f_i continuously increases along the trajectory of (Σ^*). Therefore $A_{\theta_i(0)}(t) = f_i\left(\gamma_{\theta_i(0)}(t)\right)$ is an increasing function and therefore invertible. Let us call $A^{-1}_{\boldsymbol{\theta}_i(0)}\left(f_i\right)$ the inverse. Let's recall now the $\boldsymbol{\theta}_i$ part of the dynamical system of interest Equation (1)

$$\dot{\boldsymbol{\theta}}_i = -\nabla_{\boldsymbol{\theta}_i} f_i\left(\boldsymbol{\theta}_i\right) \frac{\partial L}{\partial f_i}(\boldsymbol{F}(\boldsymbol{\theta}), \boldsymbol{G}(\boldsymbol{\phi}))$$

initialized at $\boldsymbol{\theta}_i(0)$. Applying Lemma 1 for the first equation with

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

we have that under the GDA dynamics:

$$\boldsymbol{\theta}_i(t) = \gamma_{\boldsymbol{\theta}_i(0)}\left(\int_0^t h(s)\mathrm{d}s\right) \quad (P)$$

Thus it holds

$$f_i\left(\boldsymbol{\theta}_i(t)\right) = f\left(\gamma_{\boldsymbol{\theta}_i(0)}\left(\int_0^t h(s)\mathrm{d}s\right)\right) = A_{\boldsymbol{\theta}_i(0)}\left(\int_0^t h(s)\mathrm{d}s\right)$$

or equivalently

$$\int_0^t h(s)\mathrm{d}s = A^{-1}_{\boldsymbol{\theta}_i(0)}\left(f_i\left(\boldsymbol{\theta}_i(t)\right)\right)$$

Plugging in back to Equation (P)

$$\boldsymbol{\theta}_i(t) = \gamma_{\boldsymbol{\theta}_i(0)} \left(A^{-1}_{\boldsymbol{\theta}_i(0)} \left(f_i \left(\boldsymbol{\theta}_i(t) \right) \right) \right)$$

Therefore we can pick

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

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 $X_{\boldsymbol{\phi}_j(0)}$. Let us now track the time derivative of $f_i \left(\boldsymbol{\theta}_i \right)$ and $g_j \left(\boldsymbol{\phi}_j \right)$

$$\dot{f}_i = \nabla_{\boldsymbol{\theta}_i} f_i \left(\boldsymbol{\theta}_i \right) \dot{\boldsymbol{\theta}}_i = \left\| \nabla_{\boldsymbol{\theta}_i} f_i \left(\boldsymbol{\theta}_i \right) \right\|^2 \frac{\partial L}{\partial f_i} (\boldsymbol{F}, \boldsymbol{G})$$
$$\dot{g}_j = \nabla_{\boldsymbol{\phi}_j} g_j \left(\boldsymbol{\phi}_j \right) \dot{\boldsymbol{\phi}}_j = \left\| \nabla_{\boldsymbol{\phi}_j} g_j \left(\boldsymbol{\phi}_j \right) \right\|^2 \frac{\partial L}{\partial g_j} (\boldsymbol{F}, \boldsymbol{G})$$

We can now replace $\boldsymbol{\theta}_i = X_{\boldsymbol{\theta}, (0)} \left(f_i \right)$ and $\boldsymbol{\phi}_j = X_{\boldsymbol{\phi}_\cdot (0)} \left(g_i \right)$ to get the equations required.

C. Proof for Lemma 9

Proof. Simple substitution gets us the following

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

For convex $L(\cdot, \boldsymbol{G})$ and concave $L(\boldsymbol{F}, \cdot)$, we have according to the definition:

$$\begin{aligned} - \left\langle \boldsymbol{F} - \boldsymbol{p}, \nabla_{\boldsymbol{F}} L(\boldsymbol{F}, \boldsymbol{G}) \right\rangle &\leq L(\boldsymbol{p}, \boldsymbol{G}) - L(\boldsymbol{F}, \boldsymbol{G}) \\ \left\langle \boldsymbol{G} - \boldsymbol{q}, \nabla_{\boldsymbol{G}} L(\boldsymbol{F}, \boldsymbol{G}) \right\rangle &\leq L(\boldsymbol{F}, \boldsymbol{G}) - L(\boldsymbol{F}, \boldsymbol{q}) \end{aligned}$$

D. Proof for Theorem 10

Proof. Leveraging Lemma 2, there is a function H which is well defined in $D = \left\{ \mathrm{Im}_{f_i} \left(\boldsymbol{\theta}_i(0) \right) \right\}_{i=1}^{N} \times \left\{ \mathrm{Im}_{g_j} \left(\boldsymbol{\phi}_j(0) \right) \right\}_{j=1}^{M}$ and in this domain $\dot{H} \leq 0$. Given the safety conditions we know that $(\boldsymbol{p}, \boldsymbol{q}) \in D$. Observe that for the proposed function, it holds that $H(\boldsymbol{p}, \boldsymbol{q}) = 0$. 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 $D - \{(\boldsymbol{p}, \boldsymbol{q})\}$, $H > 0$. Applying Theorem 8 for the continuously differentiable H we have that $(\boldsymbol{p}, \boldsymbol{q})$ is stable for Equation (2).

E. Proof of Theorem 11

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

$$\frac{\mathrm{d}\sigma(x)}{\mathrm{d}x} = \sigma(x)(1 - \sigma(x)).$$

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

$$(T) := \left\{ \begin{array}{rl} \dot{f}_i = & -f_i^2 \left(1 - f_i \right)^2 \frac{\partial L}{\partial f_i} (\boldsymbol{F}, \boldsymbol{G}) \\ \dot{g}_j = & g_j^2 \left(1 - g_j \right)^2 \frac{\partial L}{\partial g_j} (\boldsymbol{F}, \boldsymbol{G}) \end{array} \right\}$$

Notice that

1. The dynamical system (T) in the operator space is independent of the initial conditions. In fact, the dynamical system of (T) and the one of Equation (1), called (Σ) for short, are diffeomorphic for all initializations, not just a specific trajectory.

2. Since $(\boldsymbol{\theta}(0), \boldsymbol{\phi}(0))$ is safe, using Theorem 2 we get that $(\boldsymbol{p}, \boldsymbol{q})$ is stable for (T). We would like to prove that for every open neighborhood V of $\left(\boldsymbol{F}^{-1}(\boldsymbol{p}), \boldsymbol{G}^{-1}(\boldsymbol{q})\right)$ there exists an open neighborhood U of $\left(\boldsymbol{F}^{-1}(\boldsymbol{p}), \boldsymbol{G}^{-1}(\boldsymbol{q})\right)$ such that

$$(\boldsymbol{\theta}_{\text{init}}\,, \boldsymbol{\phi}_{\text{init}}\,) \in U \Longrightarrow \forall t \geq 0 : (\boldsymbol{\theta}(t), \boldsymbol{\phi}(t)) \in V.$$

Applying the diffeomorphism $\gamma = \gamma_{\Sigma \to T}$ between GDA dynamics of (Σ) and (T), $\gamma(V)$ is an open neighborhood of $(\boldsymbol{p}, \boldsymbol{q})$ since V is open and $\gamma\left(\left(\boldsymbol{F}^{-1}(\boldsymbol{p}), \boldsymbol{G}^{-1}(\boldsymbol{q})\right)\right) \equiv (\boldsymbol{p}, \boldsymbol{q}) \in \gamma(V)$. By Item 2 , since $(\boldsymbol{p}, \boldsymbol{q})$ is stable for (T) there is an open neighborhood $\tilde{U}$ of $(\boldsymbol{p}, \boldsymbol{q})$ such that:

$$(\boldsymbol{F}_{\text{init}}\,, \boldsymbol{G}_{\text{init}}\,) \in \tilde{U} \Longrightarrow \forall t \geq 0 : (\boldsymbol{F}(t), \boldsymbol{G}(t)) \in \gamma(V)$$

or equivalently

$$\gamma\,(\boldsymbol{\theta}_{\text{init}}\,, \boldsymbol{\phi}_{\text{init}}\,) \in \tilde{U} \Longrightarrow \forall t \geq 0 : \gamma(\boldsymbol{\theta}(t), \boldsymbol{\phi}(t)) \in \gamma(V)$$

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

$$(\boldsymbol{\theta}_{\text{init}}\,, \boldsymbol{\phi}_{\text{init}}\,) \in U \Longrightarrow \forall t \geq 0 : (\boldsymbol{\theta}(t), \boldsymbol{\phi}(t)) \in V$$

F. Proof of Theorem 12

Proof. Let us define the following sets

$$\begin{aligned}
&\forall i \in [n] : A_i = \left\{ \ \boldsymbol{\theta}_i \in \mathbb{R}^{n_i} \right. \\
&\forall j \in [m] : B_j = \left\{ \begin{array}{ll} f_i\left(\boldsymbol{\theta}_i\right) \in [p_i - \xi, p_i + \xi]\} & g_j\left(\boldsymbol{\phi}_j\right) \in [q_j - \xi, q_j + \xi] \\ \boldsymbol{\phi}_j \in \mathbb{R}^{m_j} & \end{array} \right\}
\end{aligned}$$

Since f_i and g_j are proper A_i and B_j are compact sets. Thus, the continuous functions $\left\|\nabla f_i\left(\boldsymbol{\theta}_i\right)\right\|^2$ and $\left\|\nabla g_j\left(\boldsymbol{\phi}_j\right)\right\|^2$ 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 κ_{f_i} and κ_{g_j} the minima. Observe that the minima and maxima must be all greater than zero since $[p_i - \xi, p_i + \xi]$ and $[q_j - \xi, q_j + \xi]$ are regular values. Let us define

$$\begin{aligned}
\kappa &= \min\left\{\min_{1 \leq i \leq n} \kappa_{f_i}, \min_{1 \leq j \leq m} \kappa_{g_j}\right\} \\
K &= \max\left\{\max_{1 \leq i \leq n} K_{f_i,}, \max_{1 \leq j \leq m} K_{g_j}\right\}
\end{aligned}$$

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

$$S = \left\{(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \mathbb{R}^N \times \mathbb{R}^M \quad | \quad \forall i \in [n] : \boldsymbol{\theta}_i \in A_i, \quad \forall j \in [m] : \boldsymbol{\phi}_j \in B_j\right\}$$

We can prove that every $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in S$ is a safe initialization for $(\boldsymbol{p}, \boldsymbol{q})$. Of course, every $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_j$ are not stationary points of f_i and g_j respectively. We also need to prove that the equilibrium $(\boldsymbol{p}, \boldsymbol{q})$ is feasible. We will prove this by contradiction. Let there be a $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in S$ such that $(\boldsymbol{p}, \boldsymbol{q})$ is not feasible. Without loss of generality we can assume that there is an $i \in [n]$ such that $p_i \notin \operatorname{Im}_{f_i}\left(\boldsymbol{\theta}_i\right)$. The case for the g_j is symmetrical. Along the gradient ascent trajectory of f_i with initialization at $\boldsymbol{\theta}_i$, observe that $f_i(t)$ cannot attain an infimum or a supremum in $[p_i - \xi, p_i + \xi]$ because there are no stationary points of f_i in A_i. Observe also that at initialization

$f_i(\boldsymbol{\theta}_i) \in [p_i - \xi, p_i + \xi]$. Thus Let us pick an initialization $(\boldsymbol{\theta}(0), \phi(0))$ such that $r(0) \leq \xi^2$. It is clear that $(\theta(0), \phi(0)) \in S$ and so it is safe for $(\boldsymbol{p}, \boldsymbol{q})$. We can do the same steps as in Theorem 2 to prove that the function $H(\boldsymbol{F}, \boldsymbol{G})$ below does not increase under the dynamics of Equation (1):

$$H(\boldsymbol{F}, \boldsymbol{G}) = \sum_{i=1}^{N} \int_{p_i}^{f_i} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z + \sum_{j=1}^{M} \int_{q_j}^{g_j} \frac{z - q_j}{\left\|\nabla g_j\left(X_{\boldsymbol{\phi}_j(0)}(z)\right)\right\|^2} \, \mathrm{d}z$$

Observe that since $(\boldsymbol{\theta}(0), \boldsymbol{\phi}(0)) \in S$ we have that the interval between p_i and $f_i(\boldsymbol{\theta}_i(0))$ belongs in $[p_i - \xi, p_i + \xi]$ and $\|\nabla f_i(\cdot)\|^2 \geq \kappa$ in this interval. Thus we can write

$$\frac{\left(f_i(\boldsymbol{\theta}_i(0)) - p_i\right)^2}{2\kappa} \geq \int_{p_i}^{f_i(\boldsymbol{\theta}_i(0))} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z$$

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

$$\frac{r(0)}{2\kappa} \geq H(\boldsymbol{F}(\boldsymbol{\theta}(0)), \boldsymbol{G}(\boldsymbol{\phi}(0))) \geq H(\boldsymbol{F}(\boldsymbol{\theta}(t)), \boldsymbol{G}(\boldsymbol{\phi}(t)))$$

Let us pick $r(0) < \min\left\{\xi^2, \xi^2 \frac{\kappa}{K}\right\} = \xi^2 \frac{\kappa}{K}$. 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 escaped S, then without loss of generality this means that there is at least an $i \in [n]$ such that at some $t > 0$, $f_i(\boldsymbol{\theta}_i(t)) \notin [p_i - \xi, p_i + \xi]$. The case of g_j is similar. Clearly we have that

$$\int_{p_i}^{f_i(\boldsymbol{\theta}_i(t))} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\theta_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z \geq \min\left\{\int_{p_i}^{p_i - \xi} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z, \int_{p_i}^{p_i + \xi} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z\right\}$$

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

$$\int_{p_i}^{f_i(\boldsymbol{\theta}_i(t))} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z \geq \frac{\xi^2}{2K}.$$

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

$$\frac{r(0)}{2\kappa} \geq H(\boldsymbol{F}(\boldsymbol{\theta}(t)), \boldsymbol{G}(\boldsymbol{\phi}(t))) \geq \int_{p_i}^{f_i(\boldsymbol{\theta}_i(t))} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z \geq \frac{\xi^2}{2K}.$$

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

$$\int_{p_i}^{f_i(\boldsymbol{\theta}_i(t))} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2} \, \mathrm{d}z \geq \frac{\left(f_i(\boldsymbol{\theta}_i(t)) - p_i\right)^2}{2K}.$$

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

$$\frac{r(0)}{2\kappa} \geq H(\boldsymbol{F}(\boldsymbol{\theta}(t)), \boldsymbol{G}(\boldsymbol{\phi}(t))) \geq \frac{r(t)}{2K}.$$

For every $\epsilon > 0$, there is a positive $\delta = \frac{\min\left\{\zeta^2, \epsilon\right\}\kappa}{K}$ such that

$$r(0) < \delta \Longrightarrow r(t) < \epsilon.$$

G. Proof of Lemma 13

Proof. Pick a point $(\boldsymbol{p}, \boldsymbol{q}) \in Z$. Since our initialization is safe for this saddle point, we can construct the H function as in Theorem 5 and prove that it has the following property

$$\dot{H} \leq 0 \text{ in } D = \{\text{Im}_{f_i}(\boldsymbol{\theta}_i(0))\}_{i=1}^{N} \times \left\{\text{Im}_{g_j}\left(\boldsymbol{\phi}_j(0)\right)\right\}_{j=1}^{M}$$

If $(\boldsymbol{F}(\boldsymbol{\theta}(0)), \boldsymbol{G}(\boldsymbol{\phi}(0))) = (\boldsymbol{p}, \boldsymbol{q})$ 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.

$$H_0 = \min_{(\boldsymbol{F},\boldsymbol{G}) \in \partial B} H(\boldsymbol{F}, \boldsymbol{G})$$
$$\Omega = \{(\boldsymbol{F}, \boldsymbol{G}) \in B \mid H(\boldsymbol{F}, \boldsymbol{G}) \leq H_0/2\}$$

We know that in both of the cases $H_0 > 0$ from Theorem 2 .

Since $\dot{H} \leq 0$, starting in Ω, it implies that $H(\boldsymbol{F}(t), \boldsymbol{G}(t)) \leq H_0$ for $t \geq 0$, 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 9

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

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

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

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

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

$$\dot{f}_i = 0 \Rightarrow \quad \forall i \in [N] : \quad \left\|\nabla_{\boldsymbol{\theta}_i} f_i\left(X_{\boldsymbol{\theta}_i(0)}(p_i)\right)\right\|^2 \frac{\partial L}{\partial f_i}(\boldsymbol{p}, \boldsymbol{G}) = 0$$
$$L(\boldsymbol{p}, \dot{\boldsymbol{G}}(t)) = 0 \Rightarrow \quad \sum_{j=1}^{M} \left\|\nabla_{\boldsymbol{\phi}_j} g_j\left(X_{\boldsymbol{\phi}_j(0)}(g_j)\right)\right\|^2 \left[\frac{\partial L}{\partial g_j}(\boldsymbol{p}, \boldsymbol{G})\right]^2 = 0$$

We know that $\left\|\nabla_{\boldsymbol{\theta}_i} f_i\left(X_{\boldsymbol{\theta}_i(0)}(p_i)\right)\right\| \neq 0$ by the safety conditions and that $\left\|\nabla_{\boldsymbol{\phi}_j} g_j\left(X_{\boldsymbol{\phi}_j(0)}(g_j)\right)\right\|^2 \neq 0$ inside D again by safety conditions. This implies

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

Thus $\mathcal{M}$ contains only stationary points of L so $\mathcal{M} \subseteq \text{Solution}(L)$. In addition $\mathcal{M} \subseteq D$ so only stationary points of L for which the initialization is safe are allowed so $\mathcal{M} \subseteq Z$. Applying Theorem 9 we have that for any initialization of Equation (3) inside Ω, as $t \to \infty (\boldsymbol{F}(t), \boldsymbol{G}(t))$ approaches $\mathcal{M}$ and thus Z is locally asymptotically stable for Equation (3).

H. Proof of Theorem 14

Proof. Again let's pick a point $(\boldsymbol{p}, \boldsymbol{q}) \in Z$. Since our initialization is safe for this saddle point, we can construct the H function as in Theorem 5 and prove that it has the following property

$$\dot{H} \leq 0 \text{ in } D = \left\{\text{Im}_{f_i}\left(\boldsymbol{\theta}_i(0)\right)\right\}_{i=1}^{N} \times \left\{\text{Im}_{g_j}\left(\boldsymbol{\phi}_j(0)\right)\right\}_{j=1}^{M}$$

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

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

where we know that $H_0 > 0$ from Theorem 2. Let us assume that indeed Ω is in the interior of D. Then, applying the same argumentation as in Lemma 3 combined with Theorem 2, all fixed points in Z are stable. So applying Theorem 10 we get that the trajectory initialized at $(\boldsymbol{F}(\boldsymbol{\theta}(0)), \boldsymbol{G}(\boldsymbol{\phi}(0))) \in \Omega$ 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 $(\boldsymbol{F}, \boldsymbol{G})$ 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, Ω should have no points close to the boundary of H and thus it should be in the interior.

As $(\boldsymbol{F}, \boldsymbol{G})$ approach the boundary of D, at least one of the variables f_i or g_j approaches the endpoints points of $\text{Im}_{f_i}\left(\boldsymbol{\theta}_i(0)\right)$ or $\text{Im}_{g_j}\left(\boldsymbol{\phi}_j(0)\right)$ 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 $\pm\infty$ 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 3 to get

$$\dot{f}_i = \left\|\nabla_{\boldsymbol{\theta}_i} f_i\left(X_{\boldsymbol{\theta}_i(0)}\left(f_i\right)\right)\right\|^2$$

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

$$a\left(f_i\right) = \int_{p_i}^{f_i} \frac{1}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2}\,\mathrm{d}z$$

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

$$\lim_{t\to\infty} \int_{p_i}^{f_i(t)} \frac{1}{\left\|\nabla f_i\left(X_{\theta_i(0)}(z)\right)\right\|^2}\,\mathrm{d}z = \int_{p_i}^{f_{is}} \frac{1}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2}\,\mathrm{d}z = \infty$$

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

$$f_i \to f_{is} \Longrightarrow \int_{p_i}^{f_i} \frac{z - p_i}{\left\|\nabla f_i\left(X_{\theta_i(0)}(z)\right)\right\|^2}\,\mathrm{d}z \to \infty$$

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 $(\boldsymbol{F}, \boldsymbol{G})$ 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 nonnegative so no matter what the other variables in $(\boldsymbol{F}, \boldsymbol{G})$ are doing they cannot stop $H \to \infty$.

I. Proof of Theorem 16

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

$$\frac{\partial L(\boldsymbol{x}, \boldsymbol{y})}{\partial x_i} + \lambda x_i = 0$$
$$\frac{\partial L(\boldsymbol{x}, \boldsymbol{y})}{\partial y_j} - \lambda y_j = 0$$

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

$$\mathrm{D}_{(\boldsymbol{x},\boldsymbol{y})} r(0, \boldsymbol{x}^*, \boldsymbol{y}^*) = \nabla^2 L(\boldsymbol{x}^*, \boldsymbol{y}^*)$$

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 λ and returns $g(\lambda) = (\boldsymbol{x}(\lambda), \boldsymbol{y}(\lambda))$ such that $r(\lambda, g(\lambda)) = \mathbf{0}$. Thus for a small enough λ, we have that g returns the corresponding equilibria of L'. By continuity of g, for all ϵ there is a $\delta > 0$

$$\forall 0 < \lambda < \delta : \|\boldsymbol{x}(\lambda) - \boldsymbol{x}(0)\|^2 + \|\boldsymbol{y}(\lambda) - \boldsymbol{y}(0)\|^2 \leq \epsilon^2$$

But $(\boldsymbol{x}(0), \boldsymbol{y}(0)) = (\boldsymbol{x}^*, \boldsymbol{y}^*)$ so the equilbrium of L' has an ϵ-close equilibrium of L for $\lambda < \delta$. By strict convexity strict concavity of L', it has a unique equilibrium as well. So the equilibria of L' and L are ϵ-close to each other.

J. Proof of Theorem 17

Proof. Following the same analysis with the strict convex concave analysis of the previous section, if $(\boldsymbol{F}(\boldsymbol{\theta}(0)), \boldsymbol{G}(\boldsymbol{\phi}(0))) = (\boldsymbol{p}, \boldsymbol{q})$ then the theorem holds trivially. Otherwise, since our initialization is safe for $(\boldsymbol{p}, \boldsymbol{q})$, we can construct the H function as in Theorem 2 and prove that it has the following property in $D = \{\mathrm{Im}_{f_i}(\boldsymbol{\theta}_i(0))\}_{i=1}^N \times \left\{\mathrm{Im}_{g_j}\left(\boldsymbol{\phi}_j(0)\right)\right\}_{j=1}^M$

$$\begin{aligned}
\dot{H} &\leq L'(\boldsymbol{p}, \boldsymbol{G}) - L'(\boldsymbol{p}, \boldsymbol{q}) + L'(\boldsymbol{p}, \boldsymbol{q}) - L'(\boldsymbol{F}, \boldsymbol{q}) \\
&\leq -\frac{\lambda}{2}\left(\|\boldsymbol{F}(\boldsymbol{\theta}(t)) - \boldsymbol{p}\|^2 + \|\boldsymbol{G}(\boldsymbol{\phi}(t)) - \boldsymbol{q}\|^2\right) \\
&\leq -\frac{\lambda}{2} r(t)
\end{aligned}$$

Where the second step follows from $L'(\boldsymbol{p}, \cdot)$ being λ strongly concave and $L'(\cdot, \boldsymbol{q})$ being λ strongly convex and $\boldsymbol{q}, \boldsymbol{p}$ being the corresponding optima of these functions since $(\boldsymbol{p}, \boldsymbol{q})$ is an equilibrium. Let us define

$$\begin{aligned}
H_0 &= H(\boldsymbol{F}(\boldsymbol{\theta}(0)), \boldsymbol{G}(\boldsymbol{\phi}(0))) \\
\Omega &= \{(\boldsymbol{F}, \boldsymbol{G}) \in D \mid H(\boldsymbol{F}, \boldsymbol{G}) \leq H_0\}
\end{aligned}$$

where we know that $H_0 > 0$ from Theorem 2. Additionally, we can apply Claim 1 even in the new dynamics, so Ω is in the interior of D. Since $\dot{H} \leq 0$, starting in Ω, it implies that $H(\boldsymbol{F}(\boldsymbol{\theta}(t)), \boldsymbol{G}(\boldsymbol{\phi}(t))) \leq H_0$ for $t \geq 0$, so $(\boldsymbol{F}(t), \boldsymbol{G}(t))$ stays in Ω. Additionally, Ω is closed since it is a sublevel set of a continuous function. Notice that the restriction of Ω on D does not affect the above properties since Ω is in the interior of D. Thus Ω is a compact forward invariant set. For a safe initialization $(\theta(0), \boldsymbol{\phi}(0))$, the following continuous functions must have a minimum and maximum value on Ω respectively.

$$\begin{aligned}
K_{f_i} &\geq \left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(\cdot)\right)\right\|^2 \geq \kappa_{f_i} \\
K_{g_j} &\geq \left\|\nabla g_j\left(X_{\phi_j(0)}(\cdot)\right)\right\|^2 \geq \kappa_{g_j}
\end{aligned}$$

Observe that the minima and maxima must be all greater than zero, since both $\left\|\nabla_{\boldsymbol{\phi}_j} g_j\left(X_{\boldsymbol{\phi}_j(0)}(g(t))\right)\right\|$, $\left\|\nabla_{\boldsymbol{\phi}_j} f_i\left(X_{\boldsymbol{\phi}_i(0)}(f(t))\right)\right\|$ go to 0 as this happens only at the boundaries of D which are outside Ω. Let us define

$$\begin{aligned}
\kappa &= \min\left\{\min_{1\leq i\leq n} \kappa_{f_i}, \min_{1\leq j\leq m} \kappa_{g_j}\right\} \\
K &= \max\left\{\max_{1\leq i\leq n} K_{f_{i'}} \max_{1\leq j\leq m} K_{g_j}\right\}
\end{aligned}$$

Observe that $K \geq \left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(\cdot)\right)\right\|^2 \geq \kappa$ in this interval. Thus we can write

$$\frac{\left(f_i\left(\boldsymbol{\theta}_i(t)\right)-p_i\right)^2}{2\kappa} \geq \int_{p_i}^{f_i(\boldsymbol{\theta}_i(t))} \frac{z-p_i}{\left\|\nabla f_i\left(X_{\boldsymbol{\theta}_i(0)}(z)\right)\right\|^2}\,\mathrm{d}z \geq \frac{\left(f_i\left(\boldsymbol{\theta}_i(t)\right)-p_i\right)^2}{2K}$$

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

$$\frac{r(t)}{2\kappa} \geq H(\boldsymbol{F}(\boldsymbol{\theta}(t)), \boldsymbol{G}(\boldsymbol{\phi}(t))) \geq \frac{r(t)}{2K}$$

Thus we can extend our analysis

$$\dot{H} \leq -\lambda r(t) \leq -\frac{2\kappa\lambda}{2} H(t) \Rightarrow H(t) \leq H_0 e^{-\lambda\kappa t} \Rightarrow r(t) \leq 2 \times K \times H_0 e^{-\lambda\kappa t}$$