STAY-ON-THE-RIDGE: GUARANTEED CONVERGENCE TO LOCAL MINIMAX EQUILIBRIUM IN NONCONVEX-NONCONCAVE GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Min-max optimization problems involving nonconvex-nonconcave objectives have found important applications in adversarial training and other multi-agent learning settings. Yet, no known gradient descent-based method is guaranteed to converge to (even local notions of) min-max equilibrium in the nonconvex-nonconcave setting. For all known methods, there exist relatively simple objectives for which they cycle or exhibit other undesirable behavior different from converging to a point, let alone to some game-theoretically meaningful one Vlatakis-Gkaragkounis et al. (2019); Hsieh et al. (2021). The only known convergence guarantees hold under the strong assumption that the initialization is very close to a local min-max equilibrium Wang et al. (2019). Moreover, the afore-described challenges are not just theoretical curiosities. All known methods are unstable in practice, even in simple settings.

We propose the first method that is guaranteed to converge to a local min-max equilibrium for smooth nonconvex-nonconcave objectives. Our method is second-order and provably escapes limit cycles as long as it is initialized at an easy-to-find initial point. Both the definition of our method and its convergence analysis are motivated by the topological nature of the problem. In particular, our method is not designed to decrease some potential function, such as the distance of its iterate from the set of local min-max equilibria or the projected gradient of the objective, but is designed to satisfy a topological property that guarantees the avoidance of cycles and implies its convergence.

1 INTRODUCTION

Min-max optimization lies at the foundations of Game Theory von Neumann (1928), Convex Optimization Dantzig (1951a); Adler (2013) and Online Learning Blackwell (1956); Hannan (1957); Cesa-Bianchi & Lugosi (2006), and has found many applications in theoretical and applied fields including, more recently, in adversarial training and other multi-agent learning problems Goodfellow et al. (2014); Madry et al. (2018); Zhang et al. (2019). In its general form, it can be written as

$$\min_{\theta \in \Theta} \max_{\omega \in \Omega} f(\theta, \omega), \tag{1}$$

where Θ and Ω are convex subsets of the Euclidean space, and f is continuous.

Equation (1) can be viewed as a model of a sequential-move game wherein a player who is interested in minimizing f chooses θ first, and then a player who is interested in maximizing f chooses ω after seeing θ. Solving (1) corresponds to an equilibrium of this sequential-move game.

We may also study the simultaneous-move game with the same objective f wherein the minimizing player and the maximizing player choose θ and ω simultaneously. The Nash equilibrium of the simultaneous-move game, also called a *min-max equilibrium*, is a pair $(\theta^\star, \omega^\star) \in \Theta \times \Omega$ such that

$$f(\theta^\star, \omega^\star) \leq f(\theta, \omega^\star), \quad \text{for all } \theta \in \Theta \quad \text{and} \quad f(\theta^\star, \omega^\star) \geq f(\theta^\star, \omega), \quad \text{for all } \omega \in \Omega. \tag{2}$$

It is easy to see that a Nash equilibrium of the simultaneous-move game also constitutes a Nash equilibrium of the sequential-move game, but the converse need not be true Jin et al. (2019). Here, we focus on solving the (harder) simultaneous-move game. In particular, we study the existence of *dynamics* which converge to solutions of the simultaneous-move game, namely the existence of methods that make incremental updates to a pair (θ_t, ω_t) so as the sequence (θ_t, ω_t) converges, as $t \to \infty$, to some (θ^*, ω^*) satisfying equation 2 or some relaxation of it.

This problem has been extensively studied in the special case where Θ and Ω are convex and compact and f is convex-concave — i.e. convex in θ for all ω and concave in ω for all θ. In this case, the set of Nash equilibria of the simultaneous-move game is equal to the set of Nash equilibria of the sequential-move game, and these sets are non-empty and convex von Neumann (1928). Even in this simple setting, however, many natural dynamics surprisingly fail to converge: *gradient descent-ascent*, as well as various continuous-time versions of *follow-the-regularized-leader*, not only fail to converge to a min-max equilibrium, even for very simple objectives, but may even exhibit chaotic behavior Mertikopoulos et al. (2018); Vlatakis-Gkaragkounis et al. (2019); Hsieh et al. (2021). In order to circumvent these negative results, an extensive line of work has introduced other algorithms, such as *extragradient* Korpelevich (1976) and *optimistic gradient descent* Popov (1980), which exhibit last-iterate convergence to the set of min-max equilibria in this setting; see e.g. Daskalakis et al. (2018); Daskalakis & Panageas (2018); Mazumdar & Ratliff (2018); Rafique et al. (2018); Hamedani & Aybat (2018); Adolphs et al. (2019); Daskalakis & Panageas (2019); Liang & Stokes (2019); Gidel et al. (2019); Mokhtari et al. (2019); Abernethy et al. (2019); Golowich et al. (2020b;a); Gorbunov et al. (2022); Cai et al.. Alternatively, one may take advantage of the convexity of the problem, which implies that several no-regret learning procedures, such as online gradient descent, exhibit *average*-iterate convergence to the set of min-max equilibria Cesa-Bianchi & Lugosi (2006); Shalev-Shwartz (2012); Bubeck & Cesa-Bianchi (2012); Shalev-Shwartz & Ben-David (2014); Hazan (2016). Beyond the convex/concave setting Lin et al. (2020); Kong & Monteiro (2021); Ostrovskii et al. (2021) show that convexity with respect to one of the two players is enough to design algorithms that exhibit average-iterate convergence to min-max equilibria while Diakonikolas et al. (2021) and Pethick et al. (2022) provide convergence results for *weak Minty variational inequalities*.

Our focus in this paper is on the more general case where f is not convex-concave, i.e. it may fail to be convex in θ for all ω, or may fail to be concave in ω for all θ, or both. We call this general setting where neither convexity with respect to θ nor concavity with respect to ω is assumed, the *nonconvex-nonconcave* setting. This setting presents some substantial challenges. First, min-max equilibria are *not* guaranteed to exist, i.e. for general objectives there may be no $(\theta^\star, \omega^\star)$ satisfying equation 2; this happens even in very simple cases, e.g. when $\Theta = \Omega = [0, 1]$ and $f(\theta, \omega) = (\theta - \omega)^2$. Second, it is NP-hard to determine whether a min-max equilibrium exists Daskalakis et al. (2021) and, as is easy to see, it is also NP-hard to compute Nash equilibria of the sequential-move game (which do exist under compactness of the constraint sets). For these reasons, the optimization literature has targeted the computation of local and/or approximate solutions in this setting Daskalakis & Panageas (2018); Mazumdar & Ratliff (2018); Jin et al. (2019); Wang et al. (2019); Daskalakis et al. (2021); Mangoubi & Vishnoi (2021). This is the approach we also take in this paper, targeting the computation of (ϵ, δ)*-local min-max equilibria*, which were proposed in Daskalakis et al. (2021). These are approximate and local Nash equilibria of the simultaneous-move game, defined as feasible points $(\theta^\star, \omega^\star)$ which satisfy a relaxed and local version of equation 2, namely:

$$f(\theta^\star, \omega^\star) < f(\theta, \omega^\star) + \epsilon, \text{ for all } \theta \in \Theta \text{ such that } \|\theta - \theta^\star\| \le \delta; \tag{3}$$

$$f(\theta^\star, \omega^\star) > f(\theta^\star, \omega) - \epsilon, \text{ for all } \omega \in \Omega \text{ such that } \|\omega - \omega^\star\| \le \delta. \tag{4}$$

Besides being a natural concept of local, approximate min-max equilibrium, an attractive feature of (ϵ, δ)-local min-max equilibria is that they are guaranteed to exist when f is Λ-smooth and the locality parameter, δ, is chosen small enough in terms of the smoothness, Λ, and the approximation parameter, ϵ, namely whenever $\delta \le \sqrt{\frac{2\epsilon}{\Lambda}}$. Indeed, in this regime of parameters the (ϵ, δ)-local min-max equilibria are in correspondence with the approximate fixed points of the *Projected Gradient Descent/Ascent* dynamics. Thus, the existence of the former can be established by invoking Brouwer's fixed point theorem to establish the existence of the latter. (Theorem 5.1 of Daskalakis et al. (2020)).

There are a number of existing approaches which would be natural to use to find a solution $(\theta^\star, \omega^\star)$ satisfying equation 3 and equation 4, but all run into significant obstacles. First, the idea of averaging, which can be leveraged in the convex-concave setting to obtain provable guarantees for otherwise chaotic algorithms, such as online gradient descent, no longer works, as it critically uses Jensen's inequality which needs convexity/concavity. On the other hand, negative results abound for last-iterate convergence: Hsieh et al. (2021) show that a variety of zeroth, first, and second order methods may converge to a limit cycle, even in simple settings. Vlatakis-Gkaragkounis et al. (2019) study a particular class of nonconvex-nonconcave games and show that continuous-time gradient descent-ascent (GDA) exhibits *recurrent* behavior. Furthermore, common variants of gradient descent-ascent, such as optmistic GDA (OGDA) or extra-gradient (EG), may be unstable even in the proximity of local min-max equilibria, or converge to fixed points that are not local min-max equilibria Daskalakis & Panageas (2018); Jin et al. (2019). While there do exist algorithms, such as Follow-the-Ridge proposed by Wang et al. (2019), which provably exhibit *local convergence* to a (relaxation of) local min-max equilibrium, these algorithms do not enjoy global convergence guarantees, and no algorithm is known with guaranteed convergence to a local min-max equilibrium.

These negative theoretical results are consistent with the practical experience with min-maximization of nonconvex-nonconcave objectives, which is rife with frustration as well. A common experience is that the training dynamics of first-order methods are unstable, oscillatory or divergent, and the quality of the points encountered in the course of training can be poor; see e.g. Goodfellow (2016); Metz et al. (2016); Daskalakis et al. (2018); Mescheder et al. (2018); Daskalakis & Panageas (2018); Mazumdar & Ratliff (2018); Mertikopoulos et al. (2018); Adolphs et al. (2019). In light of the failure of essentially all non-trivial, i.e., non brute-force, algorithms to guarantee convergence, even asymptotically, to local min-max equilibria, we ask the following question: *Is there any local-search algorithm which is guaranteed to converge to a local min-max equilibrium in the nonconvex-nonconcave setting? (see Table 1)*

1.1 Our Contribution

In this work we answer the above question in the affirmative: **we propose a second-order method that is guaranteed to converge to a local min-max equilibrium (Theorem 1).**. Our algorithm, called STay-ON-the-Ridge or STON'R, has some similarity to Follow-the-Ridge or FTR, which only converges locally. STON'R is the first method guaranteed to local min-max equilibrium beyond the brute-force grid-search in the non-convex/non-concave setting. Both the structure of our algorithm and its global convergence analysis are motivated by the topological nature of the problem, as established by Daskalakis et al. (2021) who showed that the problem is equivalent to Brouwer fixed point computation. In particular, the structure and analysis of STON'R are not based on a potential function argument but on a *parity argument* (see Section 4), akin to the argument used to prove the existence of Brouwer fixed points. The main challenge of our work is to prove that there exists an algorithm that uses only *local information* of the objective function f, i.e., only its second derivative, while satisfying the topological properties that are necessary to guarantee global convergence. In order to understand the main technical contributions of our paper we need first to introduce the main steps of showing the convergence using a topological argument in Section 4. Then in Section 5.4 we provide a sketch of our proof and we highlight the technical difficulties that we face.

		convex-concave	nonconvex-concave	**nonconvex-nonconcave**
Nash Eq.	existence	**yes**$^{\square}$	**no**†	**no**†
	complexity	**poly-time**‡	**NP-hard***	**NP-hard***
	convergent dynamics	**many**‡	not applicable	not applicable
Local Nash Eq.	existence	*same as above*	**yes**$^{+}$	**yes***
	complexity	*same as above*	**poly-time**$^{+}$	**PPAD-hard***
	convergent dynamics	*same as above*	**many**$^{+}$	**This paper**

Table 1: Summary of known results for simultaneous zero-sum games with differing complexity in their objective function. (□) v. Neumann (1928) (†) e.g., the min-max game with objective function $f(\theta, \omega) = -(\theta - \omega)^2$, where $\theta \in [-1, 1]$ and $\omega \in [-1, 1]$, does not have any Nash Equilibrium. ($\star$) Daskalakis et al. (2021) (‡) e.g., Dantzig (1951b); Freund & Schapire (1997); Shalev-Shwartz (2012); Cesa-Bianchi & Lugosi (2006) (+) e.g., Lin et al. (2020); Kong & Monteiro (2021); Ostrovskii et al. (2021)

2 SOLUTION CONCEPT

We begin with formulating our problem in the more general framework of *variational inequalities*. This simplifies our definitions and notations and also makes our result applicable to more general settings such as multi-player concave games Rosen (1965).

Variational Inequalities (VI). For $K \subseteq \mathbb{R}^n$, consider a continuous map $V : K \to \mathbb{R}^n$. We say that $x \in K$ is a solution of the variational inequality $\text{VI}(V, K)$ iff: $V(x)^\top \cdot (x - y) \geq 0$ for all $y \in K$.

It is well known that finding local min-max equilibria of smooth objectives can be expressed as a non-monotone VI problem. Specifically, consider the min-max optimization problem (1), take $K = \Theta \times \Omega$ and simplify notation by using $x \in K$ to denote points $(\theta, \omega) \in K$. Call the subset of coordinates of x identified with θ the "*minimizing* coordinates" and the subset of coordinates of x identified with ω the "*maximizing* coordinates." Then define $V : K \to \mathbb{R}^n$ as follows:

$$\text{For } j \in [n]\text{: set } V_j(x) := -\frac{\partial f(x)}{\partial x_j}\text{, if } j \text{ is minimizing, and } V_j(x) := \frac{\partial f(x)}{\partial x_j}\text{, otherwise.}$$

Computing (ε, δ)-local min-max equilibria of smooth objectives, i.e. points satisfying (3) & (4), can be reduced to finding solutions to $\text{VI}(V, K)$. In fact, finding even an approximate VI solution x satisfying $V(x)^\top (x - y) \geq -\alpha, \forall y \in K$, would suffice as long as $\alpha > 0$ is small enough. For more details see Theorem 5.1 of Daskalakis et al. (2020). Hence, for the rest of the paper we focus on solving variational inequality problems. For simplicity of exposition we take our constraint set to be $K = [0, 1]^n$. In this case there is a simple characterization of the solutions to $\text{VI}(V, K)$.

Definition 1. *We call a coordinate i **satisfied** at point $x \in [0, 1]^n$ if one of the following holds:*

1. *i is **zero-satisfied** at x, i.e, $V_i(x) = 0$, or*
2. *i is **boundary-satisfied** at x, i.e, ($V_i(x) \leq 0$ and $x_i = 0$) or ($V_i(x) \geq 0$ and $x_i = 1$).*

Lemma 1 (Proof in Appendix D)**.** *x is a solution of $\text{VI}(V, [0, 1]^n)$ iff j is satisfied at x, $\forall j \in [n]$.*

Finally, in the rest of the paper we make the following assumptions for V:

$$\textbf{(Λ-Lipschitz)} \quad \|V(x) - V(y)\|_2 \leq \Lambda \cdot \|x - y\|_2, \text{ for all } x, y \in [0, 1]^n.$$
$$\textbf{(L-smooth)} \quad \|J(x) - J(y)\|_F \leq L \cdot \|x - y\|_2, \text{ for all } x, y \in [0, 1]^n.$$

where J is the Jacobian of V, and $\|A\|_F$ denotes the Frobenious norm of the matrix A.

3 STAY-ON-THE-RIDGE: HIGH-LEVEL DESCRIPTION

In this section we describe our algorithm and discuss the main design ideas leading to its convergence properties presented in Section 5. As explained in the previous section, our goal is to find a point x such that every coordinate $i \in [n]$ is satisfied at x according to the Definition 1.

Our algorithm is initialized at $x(0) = (0, \ldots, 0)$. The goal of the algorithm is to satisfy all unsatisfied coordinates one-by-one in lexicographic order (although, as we will see, coordinates may go from being satisfied to being unsatisfied in the course of the algorithm). We say that our algorithm "starts epoch i at point x" iff all coordinates $\leq i - 1$ are satisfied at x and the algorithm's immediate goal is to find a point $x' \neq x$ that satisfies all coordinates $\leq i$, namely:

Goal of epoch i, starting at point x: find $x' \neq x$ satisfying all coordinates $\leq i$.

Let us assume that, at time t, our algorithm starts epoch i at point $x(t)$. Let us also assume that, at $x(t)$, all coordinates $\leq i-1$ are zero-satisfied (see Section 5.1 for the general case), i.e., $V_j(x(t)) = 0$ for all $j \leq i - 1$. Our algorithm tries to achieve the goal of epoch i starting at $x(t)$ as follows:

- Our algorithm tries to find such a point inside the connected subset $S^i(x(t)) \subseteq [0, 1]^n$ that contains all points z satisfying the following: (a) all coordinates $\leq i - 1$ are zero-satisfied at z, and (b) for all $j \geq i + 1$, $z_j = x_j(t)$.
- Our algorithm navigates $S^i(x(t))$ in the hopes of satisfying the goal of epoch i. A natural approach is to navigate $S^i(x(t))$ is to run a continuous-time dynamics $\{z(\tau)\}_{\tau \geq 0}$ that is initialized at $z(0) = x(t)$ and moves inside $S^i(x(t))$. What are possible directions of movement so that our dynamics stay within $S^i(x(t))$? If the dynamics is at some point $z \in S^i(x(t))$, it will remain in this set if it moves, infinitessimally, in a unit direction d satisfying the following constraints:

1. $d_j = 0$, for all $j \geq i+1$; */* to guarantee (b) in the definition of* $S^i(x(t))$ **/*
2. $(\nabla V_j(z))^\top \cdot d = 0$, for all $j \in \{1, \ldots, j-1\}$. */* to guarantee (a) */*

Notice that 1 and 2 specify $n-1$ constraints on n variables. We will place mild assumptions on V so that there is a unique, up to a sign flip, unit direction satisfying these constraints (see Assumption 1). Moreover, in Definition 2 we specify a rule to choose one of the two unit directions satisfying our constraints. We denote by $D^i(z)$ the direction that our tie-breaking rule selects at z.

- With the above choices, the continuous-time dynamics $\dot{z}(\tau) = D^i(z(\tau))$, initialized at $z(0) = x(t)$, is well-defined. We follow this dynamics until the earliest time that one of the following happens:
  - (Good Event): the dynamics stops at a point $x' \neq x(t)$ where coordinate i is satisfied;
  - (Bad Event): the dynamics stops at a point x' lying on the boundary of $[0,1]^n$ (and if it were to continue it would violate the constraints).

So we have described what our algorithm does if, at time t, it starts epoch i at $x(t)$. Suppose x' is the point where our dynamics executed during epoch i terminates. If the good event happened, coordinate i is satisfied at x', and our algorithm starts epoch $i+1$ at x'. If the bad event happened, our algorithm will in fact *start epoch* $i-1$ at point x'. What does this mean? That it will run the continuous-time dynamics corresponding to epoch $i-1$ on the set $S^{i-1}(x')$ starting at x' in order to find some point $x'' \neq x'$ where all coordinates $\leq i-1$ are satisfied. It may fail to do this, in which case it will start epoch $i-2$ next. Or it may succeed, in which case, it will start epoch i, and so on so forth until (as we will show!) all coordinates will be satisfied. The high-level pseudocode of our algorithm is given in Dynamics 1.

Dynamics 1 STay-ON-the-Ridge (STON'R) — High-Level Description

```
Initially $x(0) \leftarrow (0, \ldots, 0)$, $i \leftarrow 1$, $t \leftarrow 0$.
while $x(t)$ is not a VI solution do
    Initialize epoch $i$'s continuous-time dynamics, $\dot{z}(\tau) = D^i(z(\tau))$, at $z(0) = x(t)$.
    while exit condition of this dynamics has not been reached do
        Execute $\dot{z}(\tau) = D^i(z(\tau))$ forward in time.
    end while
    Set $x(t+\tau) = z(\tau)$ for all $\tau \in [0, \tau_{\text{exit}}]$ (where $\tau_{\text{exit}}$ is time exit condition was met).
    if $x(t+\tau_{\text{exit}}) \neq x(t)$ and coordinate $i$ is satisfied at $x(t+\tau_{\text{exit}})$ then
        Update the epoch $i \leftarrow i+1$.
    else
        (Bad event happened so) move to the previous epoch $i \leftarrow i-1$.
    end if
    Set $t \leftarrow t + \tau_{\text{exit}}$.
end while
return $x(t)$
```

At this point we have described an algorithm that explores the space in a natural way in its effort to satisfy coordinates, but it is unclear why it would eventually satisfy all of them, how it would escape cycles, and how it would not get stuck at non-equilibrium points. Importantly, there is no quantity that seems to be consistently improving during the execution of the algorithm.

How we can show convergence since no quantity seems to be consistently improving?

4 A TOPOLOGICAL ARGUMENT OF CONVERGENCE

Our main idea to show the convergence of the STON'R algorithm is to use a topological argument illutrated in Lemma 2 that has been employed to show the convergence of other equilibrium computation algorithms such as the elebrated Lemke-Howson algorithm Lemke & Howson (1964).

Lemma 2. *Let $G = (N, E)$ be a directed graph such that every node has in-degree at most 1 and out-degree at most 1. If there exists some node $v \in N$ with in-degree 0 and out-degree 1, then there is unique directed path starting at v and ending at some $v' \in N$ that has in-degree 1 and out-degree 0.*

The proof of Lemma 2 is straightforward, as Figure 1 illustrates. The lemma suggests a recipe for proving the convergence of some deterministic, iterative algorithm, with update rule $v_{t+1} \leftarrow F(v_t)$, whose iterates lie in a finite set N:

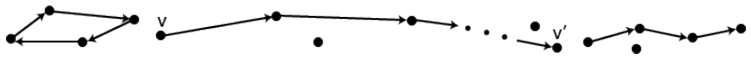


Figure 1: A directed graph whose nodes have in-degree and out-degree at most 1 is a collection of directed paths, directed cycles, and isolated nodes. Hence, if a node v has in-degree 0 and out-degree 1 then it has to be the start of a directed path that must end at a node v' after a finite number of steps.

1. Define a graph G with vertices N and edges $E = \{(u, v) \mid u \neq v \text{ and } v = F(u)\}$, i.e., there is an edge from u to v iff $v \neq u$ and v is reached after an iteration of the algorithm starting at u.
2. Argue that every vertex of G has in-degree ≤ 1. It is clear that every vertex has out-degree ≤ 1.
3. Show that the algorithm can be initialized at some v_0 that has in-degree 0 and out-degree 1.
4. Employ Lemma 2 to argue that if the algorithm is initialized at v_0 it must, eventually, arrive at some node v_{end} whose out-degree is 0. Out-degree 0 means that $v_{\text{end}} = F(v_{\text{end}})$.
5. The above prove that if the algorithm starts at v_0 it is guaranteed to converge.

In the course of the description of the algorithm and its convergence proof in Section 5, we specify a finite set of nodes N of the graph that we will construct to employ the above convergence argument. Intuitively, these are all the points at which our algorithm can possibly start a new epoch. The map $F(\cdot)$ that we use to construct our graph is the outcome of the continuous-time process that our algorithm execute when it starts an epoch at such a point.

5 DETAILED DESCRIPTION OF STON'R AND MAIN RESULT

We provide a formal description of our algorithm (Section 5.1), state our main convergence theorem (Section 5.2), and the main components of its proof building on the ideas (Section 5.4).

5.1 STON'R: DETAILED DESCRIPTION

In Section 3 we focused on the epochs where all coordinates $\leq i-1$ are zero-satisfied at the initial point x and the goal is to identify some $x' \neq x$ all coordinates $\leq i$ are satisfied. To achieve this, we execute a continuous-time dynamics constrained by keeping all coordinates $\leq i-1$ zero-satisfied. However, in the course of these dynamics we be hit the boundary. So, when we start a new epoch, some coordinates will be zero-satisfied and some will be boundary-satisfied. In that general case, the algorithm needs to execute a continuous-time dynamics constrained by keeping the zero-satisfied coordinates zero-satisfied as well as the boundary-satisfied coordinates at the right boundary.

Namely, the epochs are indexed by some coordinate $i \in [n]$ and a subset of coordinates $S \subseteq [i-1]$ that are zero-satisfied at the point x where the epoch starts. The goal of each epoch is the following.

Goal of epoch (i, S)*, starting at point* x *(where* $S \subseteq [i-1]$*, coordinates in* S *are zero-satisfied and coordinates in* $[i-1] \setminus S$ *are boundary-satisfied): find* $x' \neq x$ *where all coordinates* $\leq i$ *are satisfied, all coordinates in* S *are zero-satisfied, and all coordinates in* $[i-1] \setminus S$ *are boundary-satisfied.*

Epoch (i, S) starting at x might achieve its goal or end before it achieves its goal. In both cases, a new epoch will start. Within each epoch our algorithm executes a continuous-time dynamics that maintains all the coordinates $j \in S$ zero-satisfied, all the coordinates $j \in [i-1] \setminus S$ boundary-satisfied, and leaves all coordinates $[n] \setminus [i]$ unchanged.

Definition 2 (Tangent Unit Vector of Epoch (i, S))**.** *Let* $i \in [n]$*,* $S = \{s_1, \ldots, s_m\} \subseteq [i-1]$*, and* $x \in [0,1]^n$*, we say that a unit vector* $d \in \mathbb{R}^n$ *is* admissible *if:*

1. $d_j = 0$*, for all* $j \notin S \cup \{i\}$*, and*
2. $\nabla V_j(x)^\top \cdot d = 0$*, for all* $j \in S$*, and*
3. *the sign of* $\begin{vmatrix} \frac{\partial V_{s_1}(x)}{\partial x_{s_1}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_1}} & \cdots & \frac{\partial V_{s_m}(x)}{\partial x_{s_1}} & d_{s_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial V_{s_1}(x)}{\partial x_{s_m}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_m}} & \cdots & \frac{\partial V_{s_m}(x)}{\partial x_{s_m}} & d_{s_m} \\ \frac{\partial V_{s_1}(x)}{\partial x_i} & \frac{\partial V_{s_2}(x)}{\partial x_i} & \cdots & \frac{\partial V_{s_m}(x)}{\partial x_i} & d_i \end{vmatrix}$ *equals the sign of* $(-1)^{|S|}$.

If there is a unique unit direction satisfying the above constraints, we denote that direction $D_S^i(x)$.

Conditions 1 and 2 above describe a line in $\mathbb{R}^n$ and condition 3 specifies a direction on this line. We will place some assumptions on V so that $D^i_S(x)$ is defined for all $x \in [0,1]^n$ where coordinates S are zero-satisfied (see Assumption 1). Now, when we start epoch (i, S) at point x, we will execute the continuous-time dynamics $\dot{z}(\tau) = D^i_S(z(\tau))$, initialized at $z(0) = x$, forward in time. We this dynamics until the earliest time τ_{exit} such that $z(\tau_{\text{exit}})$ is an *exit point* according to the next definition.

Definition 3. *Suppose $i \in [n]$, $S \subseteq [i-1]$, at $x' \in [0,1]^n$, the coordinates in S are zero-satisfied at x', the coordinates in $[i-1] \setminus S$ are boundary-satisfied at x'. Then x' is an* exit point *for epoch (i, S) iff it satisfies one of the following:*

- ***(Good Exit Point)**: Coordinate i is satisfied at x', i.e., $V_i(x') = 0$, or $x'_i = 0$ and $V_i(x') < 0$, or $x'_i = 1$ and $V_i(x') > 0$.*
- ***(Bad Exit Point)**: $\exists j \in S \cup \{i\}$ s.t. $(D^i_S(x'))_j > 0$ and $x'_j = 1$, or $(D^i_S(x'))_j < 0$ and $x'_j = 0$, i.e., if the dynamics of epoch (i, S) were to continue from x', they would violate the constraints.*
- ***(Middling Exit Point)**: $\exists j \in [i-1] \setminus S$ s.t. $V_j(x') = 0$ and $(\nabla V_j(x')^\top D^i_S(x') > 0$ and $x'_j = 0)$ or $(\nabla V_j(x')^\top D^i_S(x') < 0$ and $x'_j = 1)$, i.e., if the dynamics for epoch (i, S) were to continue from x', some boundary-satisfied coordinate would become unsatisfied.*

We will place some assumptions on V so that there can be a unique j triggering the condition of Bad Exit Point and there can be a unique j triggering the Middling Exit Point condition (see Assumptions 2). Below we describe the actions that we take when one of the above exit conditions is triggered.

Action at Good Events. In case of a good event, we start epoch $(i+1, S')$ at x', where $S' = S \cup \{i\}$, if i is zero-satisfied at x', and $S' = S$, if i is boundary-satisfied at x'.

Action at Bad Events. In case of a bad event, note that the coordinate j responsible for the condition in the bad event must belong to $S \cup \{i\}$ because in all other coordinates $(D^i_S(x'))_j = 0$ by definition. Our action depends on which j triggers the bed event as follows:
(1) if the triggering $j = i$, then we start epoch $(i-1, S \setminus \{i-1\})$ at x', otherwise
(2) if the triggering $j \neq i$, then we start epoch $(i, S \setminus \{j\})$ at x'.

Action at Middling Events. In this case, we start epoch $(i, S \cup \{j\})$ at x' because the coordinate j is both zero- and boundary-satisfied at x' so we add j to S to keep it zero-satisfied next.

Combining the above rules we get a full description of our algorithm in Dynamics 2. In Appendix B we do a step-by-step execution of this algorithm for a simple 2D min-max optimization problem.

Dynamics 2 STay-ON-the-Ridge (STON'R)

```
Initially $x(0) \leftarrow (0, \dots, 0)$, $i \leftarrow 1$, $S \leftarrow \emptyset$, $t \leftarrow 0$.
while $x(t)$ is not a VI solution do
    Initialize epoch $(i, S)$'s continuous-time dynamics, $\dot{z}(\tau) = D^i_S(z(\tau))$, at $z(0) = x(t)$.
    while $z(\tau)$ is not an exit point as per Definition 3 do
        Execute $\dot{z}(\tau) = D^i_S(z(\tau))$ forward in time.
    end while
    Set $x(t+\tau) = z(\tau)$ for all $\tau \in [0, \tau_{\text{exit}}]$ (where $\tau_{\text{exit}}$ is the time $z(\tau)$ became an exit point).
    if $x(t+\tau_{\text{exit}})$ is (Good Exit Point) as in Definition 3 then
        if $i$ is zero-satisfied at $x(t+\tau_{\text{exit}})$ then
            Update $S \leftarrow S \cup \{i\}$.
        end if
        Update $i \leftarrow i+1$.
    else if $x(t+\tau_{\text{exit}})$ is a (Bad Exit Point) as in Definition 3 for $j = i$ then
        Update $i \leftarrow i-1$ and $S \leftarrow S \setminus \{i-1\}$.
    else if $x(t+\tau_{\text{exit}})$ is a (Bad Exit Point) as in Definition 3 for $j \neq i$ then
        Update $S \leftarrow S \setminus \{j\}$.
    else if $x(t+\tau_{\text{exit}})$ is a (Middling Exit Point) as in Definition 3 for $j < i$ then
        Update $S \leftarrow S \cup \{j\}$.
    end if
    Set $t \leftarrow t + \tau_{\text{exit}}$.
end while
return $x(t)$
```

5.2 OUR ASSUMPTIONS AND OUR MAIN THEOREM

We next present the assumptions on V that are needed for our convergence proof. We discuss these assumptions further in Appendix A where we present some high level reasons why they are mild.

Assumption 1. *There exist real numbers* $\sigma_{\max} > \sigma_{\min} > 0$ *such that: for all* $x \in [0,1]^n$ *and for all* $S = \{s_1, \ldots, s_m\} \subseteq [n]$, *if* $V_\ell(x) = 0$ *for all* $\ell \in S$, *then the singular values of the* $m \times m$ *matrix* $J_S^K(x)$ *are greater than* σ_{min} *and less than* σ_{max}, *where*

$$J_S^K(x) := \begin{pmatrix} \frac{\partial V_{s_1}(x)}{\partial x_{s_1}} & \ldots & \frac{\partial V_{s_1}(x)}{\partial x_{s_m}} \\ \vdots & & \vdots \\ \frac{\partial V_{s_m}(x)}{\partial x_{s_1}} & \ldots & \frac{\partial V_{s_m}(x)}{\partial x_{s_m}} \end{pmatrix}.$$

Assumption 1 ensures that the direction $D_S^i(\cdot)$ of Definition 2 is uniquely defined (see Lemma 11 in Appendix 8).

Assumption 2. *For all* $x \in [0,1]^n$, *for all* $i \in [n]$, *and for all* $S \subseteq [i-1]$: *if* $V_\ell(x) = 0 \ \forall \ell \in S$ *and* $x_\ell \in \{0,1\} \ \forall \ell \notin S \cup \{i\}$ *then there is at most one coordinate* $j \in S \cup \{i\}$ *such that* $x_j \in \{0,1\}$.

Assumption 2 ensures that any time at most one coordinate can trigger a middling or a bad event. To see this, imagine there are two different coordinates j_1, j_2 triggering a bad event at x, then $x_{j_1} \in \{0,1\}, x_{j_2} \in \{0,1\}$ and $V_{j_1}(x) = V_{j_2}(x) = 0$ and therefore Assumption 2 is violated. A similar observation applies for middling events. See also Lemma 8 and Lemma 10 in the Appendix.

Assumption 3. *For all* $x \in [0,1]^n$, *for all* $i \in [n]$, *for all* $S \subseteq [i-1]$ *such that* $V_\ell(x) = 0 \ \forall \ell \in S$ *and* $x_\ell \in \{0,1\} \ \forall \ell \notin S \cup \{i\}$, *and for all vectors* $(d_{s_1}, \ldots, d_{s_m}, d_i)$ *satisfying the equations,*

$$\nabla_{S \cup \{i\}} V_j(x)^\top \cdot (d_{s_1}, \ldots, d_{s_m}, d_i) = 0 \text{ for all } j \in S,$$

we have that $d_j \neq 0$ *if* $x_j = 0$ *or* $x_j = 1$.

Assumption 3 ensures that we can determine whether a coordinate begins or stops being satisfied by looking at the Jacobian of V. For example, consider a coordinate j such that $x_j = 0$ and $V_j(x) = 0$. If also $D_S^i(x)^\top V_j(x) = 0$ then higher-order information is needed in order to determine whether the direction $D_S^i(\cdot)$ makes the coordinate j satisfied or unsatisfied (see Lemma 4 in the Appendix).

We are now ready to state our main theorem.

Theorem 1. *Under Assumptions 1, 2, and 3, there exists some* $\bar{T} = \bar{T}(\sigma_{\min}, \sigma_{\max}, n, L, \Lambda) > 0$ *such that* STAY-ON-THE-RIDGE *(Dynamics 2) will stop, at some time* $T \leq \bar{T}$, *at some point* $x(T) \in [0,1]^n$ *that is a solution of* $\mathrm{VI}(V, [0,1]^n)$.

Remark 1 (Discrete-time Algorithm)**.** *It is possible to combine the proof of Theorem 1 with standard numerical analysis techniques to show the convergence of a simple discrete version of the dynamics assuming that the step size is small enough. For more details about this we refer to Appendix J.*

5.3 SIMULATED 2-DIMENSIONAL EXAMPLE

In Figure 2 we present the behavior of the main existing algorithms for min-max optimization in the 2-dimensional min-max problem with objective $f(\theta, \omega) := -\theta\omega - \frac{1}{20} \cdot \omega^2 + \frac{2}{20} \cdot S\left(\frac{\theta^2 + \omega^2}{2}\right) \cdot \omega^2$, where $S(z)$ is the smooth-step function equal to 0 for $z \leq 0$, 1 for $z \geq 1$ and $z^2 - 2z^3$ otherwise. With **blue** we observe the behavior of GDA, EG, and OGDA that have the same behavior in this example when initialized at $(-0.5, -1)$. With **orange** we observe the behavior of the follow-the-ridge (FtR) algorithm initialized at $(-0.5, -0.5)$ and with **green** we observe the behavior of STON'R. As we can see GDA, EG, OGDA are getting trapped to a cycle whereas FtR hits the boundary at $(-1, -1)$ that does not correspond to an equilibrium point. Our algorithm is the only one that directly converges to the equilibrium following a very short path. In Appendix C we provide a more detailed explanation of this example and we observe similar behavior for different initializations of GDA, EG, OGDA, and FtR.

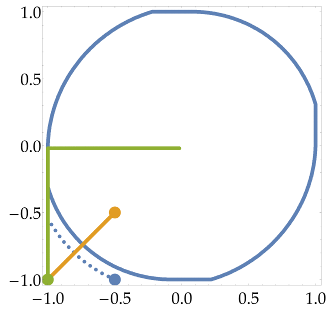


Figure 2

5.4 SKETCH OF PROOF OF THEOREM 1

For a sketch of our proof of Theorem 1 we follow the recipe that we described in Section 4. During this proof sketch we highlight some technical challenges that we face. (The full proof can be found in Appendix E.)

1. We start with the definition of the set of nodes N. The set N contains triples of the form (i, S, x) where $i \in [n]$, S is a subset of $[i-1]$ and $x \in [0,1]^n$ that satisfies the following:

   (a) all coordinates in S are zero-satisfied, (b) all coordinates in $[i-1] \setminus S$ are boundary-satisfied, (c) $x_j = 0$ for all $j \geq i+1$, and either (d1) $x_i = 0$ or (d2) x is an exit point for epoch (i, S) according to Definition 3 [1].

   Our first technical challenge is to show that the size of N is finite (see Lemma 3 in the Appendix). Next we describe a mapping $F : N \to N$. Let $(i, S, x) \in N$, we use the dynamics $\dot{z} = D^i_S(z)$ with initial condition $z(0) = x$ and we find the minimum time τ_{exit} such that $z(\tau_{\text{exit}})$ is an exit point. We then update i, S to i', S' according to the rules for actions on exit points of Section 5.1 and we define $F((i, S, x)) = (i', S', z(\tau_{\text{exit}}))$. One of our main technical challenges is to show that the dynamics $\dot{z} = D^i_S(z)$ have a unique solution under our assumptions and hence F is well defined (see Lemma 4 in the Appendix).

   The set N and the mapping F define the directed graph G, as described in Section 4, that is guaranteed to have vertices with out-degree at most 1. We also show that any $v \in V$ with out-degree 0 is an equilibrium point (see Lemma 4 in the Appendix).

2. To show that the in-degree is at most 1, we face our next technical challenge which is to show that we can actually solve the dynamics backwards in time. In particular, if we specify $z(0)$ and there is the smallest time τ_{exit} such that $z(-\tau_{\text{exit}})$ is an exit point then $z(-\tau_{\text{exit}})$ is uniquely determined. This means that there exists $F^{-1} : N \to N$ such that if $v' = F(v)$ then $F^{-1}(v') = v$ which means that no vertex in N can have in-degree more than 1 (see Lemma 5 in the Appendix).
3. We show that $v_0 = (1, \emptyset, (0, \ldots, 0)) \in N$ and that if run the dynamics $\dot{z} = D^1_{\emptyset}(z)$ backwards in time starting at $z(0) = 0$ then we get outside $[0,1]^n$ and so v_0 has in-degree 0. We also show that the dynamics $\dot{z} = D^1_{\emptyset}(z)$ can move forward in time and stay inside $[0,1]^n$ so v_0 has out-degree 1 (see Lemma 6 in the Appendix).
4. The above show that our algorithm converges according to Section 4.

6 CONCLUSIONS

Summary. In this work we propose a novel local-search algorithm, called STON'R, that is guaranteed to converge to local min-max equilibrium in the general case of non-convex non-concave objectives. To the best of our knowledge STON'R is the first method, beyond trivial brute-force, that is guaranteed to find a local min-max equilibrium starting from a simple initialization. We remark that existing min-max optimization methods required either convexity (resp. concavity) in one of the players or an initialization very close to the optimal point in order to guarantee convergence. Finally, our approach differs from existing methods in the fundamental way that both its design and analysis are based on topological rather than potential arguments. We believe that these types of arguments can play an important role in the future of multi-agent machine learning.

Comparison with Brute-Force. Since we assume that V is a Lipschitz function and that K is an n-dimensional hypercube, it is not hard to see that there exists a small enough discretization of the space such that the brute-force search over all the discrete points is guaranteed to find a solution. Such brute-force algorithms exist in most of the optimization problems like solving linear programs or finding Nash equilibria in normal form games. These trivial algorithms suffer from the curse of dimensionality even in very simple instances and hence they are almost never useful. Instead local-search algorithms such as simplex or Lemke-Howson Lemke & Howson (1964) have been extremely successful in practice because they converge very fast in the majority of real world instances although in the worst-case their complexity is the same as the brute-force. Our contribution is to provide the first such algorithm for the fundamental problem of nonconvex-nonconcave min-max optimization and we believe that it will play an important role in the future of multi-agent optimization in machine learning.

---

[1] The actual set of nodes that we used in the proof does not contain the information of i and S but we refer to the Appendix for the exact proof.

REFERENCES

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

Ilan Adler. The equivalence of linear programs and zero-sum games. *International Journal of Game Theory*, 42(1):165–177, 2013. doi: 10.1007/s00182-012-0328-8.

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 486–495, 2019.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(1):1–8, 1956. doi: 10.2140/pjm.1956.6.1.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games.

Nikolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

George B. Dantzig. A proof of the equivalence of the programming problem and the game problem. In *Koopmans, T. C., editor(s), Activity Analysis of Production and Allocation*. Wiley, New York, 1951a.

George B Dantzig. A proof of the equivalence of the programming problem and the game problem. *Activity analysis of production and allocation*, 13, 1951b.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pp. 9236–9246, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *Innovations in Theoretical Computer Science*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. *CoRR*, abs/2009.09623, 2020. URL `https://arxiv.org/abs/2009.09623`.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The Complexity of Constrained Min-Max Optimization. In *Proceedings of the 53rd ACM Symposium on Theory of Computing (STOC)*, 2021.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2746–2754. PMLR, 2021.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811, 2019.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pp. 1758–1784. PMLR, 2020b.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2014.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022.

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3: 97–139, 1957.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Arieh Iserles. *A first course in the numerical analysis of differential equations*. Cambridge university press, 2009.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

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

Carlton E Lemke and Joseph T Howson, Jr. Equilibrium points of bimatrix games. *Journal of the Society for industrial and Applied Mathematics*, 12(2):413–423, 1964.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 907–915, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Oren Mangoubi and Nisheeth K Vishnoi. Greedy adversarial equilibrium: An efficient alternative to nonconvex-nonconcave min-max optimization. In *Proceedings of the 53rd ACM Symposium on Theory of Computing (STOC)*, 2021.

Eric Mazumdar and Lillian J Ratliff. On the convergence of gradient-based learning in continuous games. *arXiv preprint arXiv:1804.05464*, 2018.

Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pp. 3481–3490, 2018.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31 (4):2508–2538, 2021.

Thomas Pethick, Puya Latafat, Panos Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *The Tenth International Conference on Learning Representations, ICLR 2022*, 2022.

L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, Nov 1980.

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

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. doi: 10.1561/2200000018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

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. *Advances in Neural Information Processing Systems*, 32, 2019.

John von Neumann. Zur Theorie der Gesellschaftsspiele. In *Math. Ann.*, pp. 295–320, 1928.

Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*, 2019.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

APPENDIX

A DISCUSSION OF ASSUMPTIONS 1, 2 AND 3

In this section we discuss about the generality of our Assumptions 1, 2, and 3. We follow a general recipe in our arguments. In particular, we consider any VI problem $\mathrm{VI}(V, K)$ that does not satisfy some of our Assumptions, then we argue that there exists a small random perturbation $\mathrm{VI}(\tilde{V}, \tilde{K})$ of $\mathrm{VI}(V, K)$ such that: (1) any approximate solution of $\mathrm{VI}(\tilde{V}, \tilde{K})$ is also an approximate solution of $\mathrm{VI}(V, K)$ with slightly higher approximation loss, and (2) $\mathrm{VI}(\tilde{V}, \tilde{K})$ satisfies all our Assumptions.

The arguments that we present in the next sections are heuristic but we conjecture that our statements are true in general which we leave as an interesting open problem. The main component that we miss towards this direction is the following: we can so that for a particular problem $\mathrm{VI}(V, K)$ if a particular point $x \in K$ violates some of the assumptions then a random perturbation suffices to make x satisfy all the assumptions. The argument is missing is to show that these random perturbations produce instances that satisfy the assumptions for every point in the space. As we said before we conjecture that this is actually true and we have verified our conjecture in some simple experiments.

A.1 ASSUMPTION 1

Let $\tilde{V}$ be a vector field such that for all $x \in K$, $\|\tilde{V}(x)\| \leq \epsilon$. If $\tilde{x}$ is an α-approximate solution to the VI problem $VI(V + \tilde{V}, K)$, i.e, $(V(x) + \tilde{V}(x))^\top (x - y) \geq -\alpha \forall y \in K$ then $\tilde{x}$ is an $\alpha + \epsilon \cdot R$-approximate solution of $VI(V, K)$, where R is the diameter of K. To see this observe that

$$\begin{aligned} V(x)^\top (x - y) &= (V(x) + \tilde{V}(x))^\top (x - y) - \tilde{V}(x)^\top (x - y) \\ &\geq -\alpha - \epsilon \cdot R. \end{aligned}$$

So our idea is that a $\tilde{V}$ with small magnitude and enough randomness should suffice to slightly perturb the problem can make sure that V and J_V do not vanish simultaneously.

In one dimension, consider the instance $V(x) = 3x^2$ which is the simplest single-dimensional VI problem violating Assumption 1. This corresponds to a local maximization problem with objective function $f(x) = x^3$ as per our discussion in Section 2. In this case, at $x = 0$ we have that $V(0) = 0$ and $V'(0) = 0$ at the same time and hence Assumption 1 is violated. However, it is easy to perturb f and V in this problem to a problem that does not have this issue. We can simply add to f a periodic function, e.g., $\alpha \cdot \sin(x + \psi)$, with parameter α very small and in particular $\alpha \leq \epsilon$ and we suppose that we chose ψ uniformly. Let $\tilde{f}$ be the modified maximization objective, i.e., $\tilde{f}(x) = f(x) + \alpha \cdot \sin(x + \psi)$.

In higher dimensions the design of the appropriate perturbation $\tilde{V}$ is more challenging and we do not have a specific construction. We conjecture though is an appropriate $\tilde{V}$ exists such that $\tilde{V}$ has small magnitude and enough randomness even in high-dimensions. To support this conjecture we ran some simple experiments with objective functions that do not satisfy 1 and we observe that indeed small random perturbations always produce objective functions that satisfy Assumption 1. A complete theoretical construction of $\tilde{V}$ is a very interesting open problem.

A.2 ASSUMPTION 2

To reason for the mildness of this assumption we restricting the domain of each variable i in the subset $[\alpha_i, 1 - \beta_i]$ of $[0, 1]$, where α_i, β_i are uniformly random in $[0, \epsilon]$ and then we apply a simple change of variables $z_i = (b_i - x_i)/(b_i - \alpha_i)$ so that the domain becomes again $[0, 1]^d$ with respect to z. If we choose α_i and β_i to be very small, a solution to the VI problem with respect to the z-variables, corresponds to a $\Theta(\epsilon)$-approximate solution with respect to to x-variables (by performing the inverse transformation). In order to exhibit why the Assumption 2 is satisfied in the perturbed domain we present the following examples.

Example 1. *Consider the function* $f(x, y) := -x^2/2 + x \cdot y + y$ *and notice that* f *does not satisfy the Assumption* 2 *at point* $(0, 0)$*. The points at which coordinate* x *is zero-satisfied belong in the curve* $C = \{(x, y) \in [0, 1]^2$ *such that* $x - y = 0\}$ *while there is no point in* $[0, 1]^2$ *at which coordinate* y *is* *satisfied. If we select* $\alpha_1, \alpha_2, \beta_1, \beta_2$ *uniformly at random in* $[0, \epsilon]$ *then*

$$\Pr[(\alpha_1, \alpha_2) \in C \textit{ or } (\alpha_1, \beta_2) \in C \textit{ or } (\beta_1, \alpha_2) \in C \textit{ or } (\beta_1, \beta_2) \in C] = 0$$

As an example, consider the curve $C = \{x \in [0, 1]^n \text{ such that } V_1(x) = 0, \ldots, V_{n-1}(x) = 0\}$ and assume (because of Assumption 1) that for all $x \in C$ the matrix

$$J(x) := \begin{pmatrix} \frac{\partial V_1(x)}{\partial x_1} & \cdots & \frac{\partial V_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial V_{n-1}(x)}{\partial x_1} & \cdots & \frac{\partial V_{n-1}(x)}{\partial x_n} \end{pmatrix}$$

admits singular values greater than σ_{min} and smaller than σ_{max}. If the boundaries $[\alpha_i, 1 - \beta_i]$ for each coordinate i are selected uniformly at random from the interval $[0, \epsilon]$, then with high probability the curve C hits the random rectangle $[\alpha_1, 1 - \beta_1] \times \cdots \times [\alpha_n, 1 - \beta_n]$ only in *pure facets* (only one coordinate i equals α_i or $1 - \beta_i$).

A.3 ASSUMPTION 3

We argue about the generality of Assumption 3 using the same idea as before. We argue that there exists a small random perturbation of every problem so that the resulting VI satisfies Assumption 3 with high probability. In particular, consider any VI problem with map $V(x)$ and define $\tilde{V}(x) = V(x) + Ax$, where each entry A_{ij} is selected uniformly at random from $[-\epsilon, \epsilon]$. A VI solution x^* for $\tilde{V}$ is a $\Theta(\epsilon n)$-approximate VI solution for V.

Now Item 2 of Definition 2 defining the notion of direction $d = D^i_S(x)$ takes the following form,

$$\left(\nabla_{S \cup \{i\}} V_j(x) + A^j_{S \cup \{i\}}\right)^\top \cdot (d_{s_1}, \ldots, d_{s_m}, d_i) = 0$$

where $A^j_{S \cup \{i\}}$ denotes the j-th row of A restricted to the columns $\ell \in S \cup \{i\}$. Due to the fact that all vectors $\nabla_{S \cup \{i\}} V_j(x)$ are linearly independent and the fact that the entries A_{ij} have been selected uniformly at random in $[-\epsilon, \epsilon]$ we can easily conclude that

$$\Pr[\text{there exists } j \in S \cup \{i\} \text{ with } d_j = 0] = 0$$

which suggests that Assumption 3 holds with high probability at x.

B 2-D EXAMPLE OF STON'R EXECUTION

In Figure 3, we show the trajectory that our algorithm follows when it is applied to solve a min-max optimization problem with objective $f(\theta, \omega) := (\theta - 1/2) \cdot (\omega - 1/2)$ where θ is the minimizing and ω is the maximizing variable. We explain below how this trajectory is derived by following Dynamics 2.

First, using our notation in Section 2, let x_1 correspond to θ and x_2 correspond to ω. As explained in the same section, finding a local min-max equilibrium can be reduced to a non-monotone VI problem where $V_1(x_1, x_2) := 1/2 - x_2$ and $V_2(x_1, x_2) := x_1 - 1/2$. Next we describe the steps that our algorithm follows.

- ▷ $x(0) = (0, 0), i = 1, S = \emptyset, t = 0$, STON'R goes to Step 3. $V_1(0, 0) = 1/2 > 0$ and $x_1 = 0$, hence coordinate 1 is not satisfied. Thus, the loop of Step 2 is activated and STON'R goes to Step 3.
- ▷ STON'R goes to Step 5 and executes $\dot{z}(\tau) = (1, 0)$ with initialization $z(0) = (0, 0)$. Note that at $x = (0, 0)$ the only unit direction satisfying the constraints of Definition 2 is $(1, 0)$ and that the same is true for any point $(\cdot, 0)$. Thus, for all these points $D^1_\emptyset((\cdot, 0)) = (1, 0)$, and the continuous-time dynamics executed at Step 5 is $\dot{z}(\tau) = (1, 0)$.
- ▷ STON'R goes to Step 7 and sets $x(1) = (1, 0)$. For any point $z = (z_1, 0)$, $V_1(z) = 1/2$. Thus the continuous-time dynamics of Step 5 only terminates when it hits the boundary of the square at point $(1, 0)$, which happens at time $\tau_{\text{exit}} = 1$. At Step 7, the algorithm sets $x(1) = z(1) = (1, 0)$.

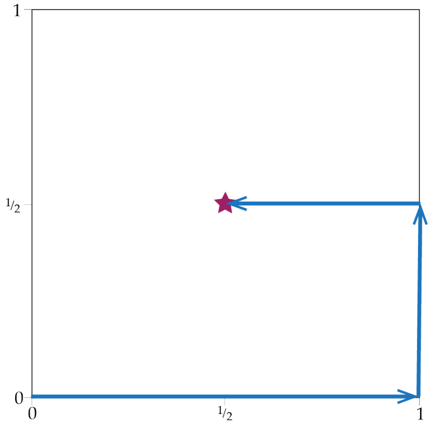


Figure 3: The path of STON'R for $f(\theta, \omega) = (\theta - 1/2) \cdot (\omega - 1/2)$.

▷ STON'R goes to Step 12 and sets $i = 2$. $V_1(x(1)) = 1/2 > 0$ thus coordinate 1 is boundary-satisfied at this point. Because this is the good event of Definition 3, the condition of the if statement of Step 8 triggers. Because coordinate 1 is boundary-satisfied the condition of the if statement of Step 9 is not triggered. Thus the algorithm arrives at Step 12 and sets $i = 2$.

▷ STON'R goes to Step 3 with $i = 2$, $S = \emptyset$. At $x(1) = (1, 0)$ coordinate 1 is boundary-satisfied since $V_1(1, 0) = 1/2 > 0$ but coordinate 2 is not satisfied since $V_2(1, 0) = 1/2 > 0$. Thus, the while condition of Step 2 is triggered and STON'R goes to Step 3.

▷ STON'R goes to Step 5 and executes $\dot{z}(\tau) = (0, 1)$ with initialization $z(0) = (1, 0)$. Note that at $x = (1, 0)$ the only unit direction satisfying the constraints of Definition 2 is $(0, 1)$ and that the same is true for any point $(1, \cdot)$. Thus, for all these points $D^2_{\emptyset}((1, \cdot)) = (0, 1)$, and the continuous-time dynamics executed at Step 5 is $\dot{z}(\tau) = (0, 1)$.

▷ STON'R goes to Step 7 and sets $x(1.5) = (1, 0.5)$. For any point $z = (1, z_2)$, $V_1(z) = 1/2 - z_2$ and $V_2 = 1/2$. Thus the continuous-time dynamics of Step 5 only terminates when it hits point $(1, 0.5)$, which happens at time $\tau_{\text{exit}} = 1/2$. The reason the continuous-time dynamics terminates at this point is because the middling condition of Definition 3 is triggered for $j = 1$. Indeed, coordinate 1 is boundary satisfied from the beginning of the continuous-time dynamics until it reaches point $(1, 0.5)$ but if the continuous-time dynamics were to continue onward, then coordinate 1 would become unsatisfied as V_1 would turn negative. Thus the continuous-time dynamics stops at time $\tau_{\text{exit}} = 1/2$, the algorithm moves to Step 7 and it sets $x(1.5) = z(0.5) = (1, 0.5)$.

▷ STON'R goes to Step 18 and sets $S = \{1\}$. Since the most recently executed continuous-time dynamics at Step 5 ended at a middling exit point, the condition of Step 17 is activated, so the algorihtm moves to Step 18 where S is set to $\{1\}$.

▷ STON'R goes to Step 3 with $i = 2$, $S = \{1\}$. At $x(1.5) = (1, 0.5)$ coordinate 1 is both zero- and boundary-satisfied since $V_1(1, 0.5) = 0$ but coordinate 2 is still not satisfied since $V_2(1, 0.5) = 1/2$. Thus, the while condition of Step 2 is triggered and STON'R goes to Step 3.

▷ STON'R goes to Step 5 and executes $\dot{z}(\tau) = (-1, 0)$ with initialization $z(0) = (1, 0.5)$. Note that at $x = (1, 0.5)$ the only unit direction satisfying the constraints of Definition 2 is $(-1, 0)$ and that the same is true for any point $(\cdot, 0.5)$. Thus, for all these points $D^2_{\{1\}}((\cdot, 0.5)) = (-1, 0)$, and the continuous-time dynamics executed at Step 5 is $\dot{z}(\tau) = (-1, 0)$.

▷ STON'R goes to Step 7 and sets $x(2) = (0.5, 0.5)$. For any point $z = (z_1, 0.5)$, $V_1(z) = 0$ and $V_2 = z_1 - 1/2$. Thus the continuous-time dynamics of Step 5 only terminates when it hits point $(0.5, 0.5)$, which happens at time $\tau_{\text{exit}} = 1/2$. The reason the continuous-time dynamics terminates at this point is because the good condition of Definition 3 is triggered for $i = 2$ at this point. Thus the continuous-time dynamics stops at time $\tau_{\text{exit}} = 1/2$, the algorithm moves to Step 7 and it sets $x(2) = z(0.5) = (0.5, 0.5)$.

- ▷ STON'R goes to Step 22 and outputs $(0.5, 0.5)$. The condition of the if statement of both Steps 8 and 9 are triggered, so $S = \{1, 2\}$ and $i = 3$. At $x(2) = (0.5, 0.5)$ both coordinate 1 and coordinate 2 are satisfied, so the while loop of Step 2 is not activated. So the algorithm goes to Step 22 and returns $(0.5, 0.5)$.

It is easy to verify that the point $(\theta, \omega) = (1/2, 1/2)$ is a (local) min-max equilibrium of $(\theta - 1/2) \cdot (\omega - 1/2)$.

C SIMULATED 2-DIMENSIONAL EXPERIMENTS

As a warm-up we present some simulated experiments to compare the performance of our algorithm with the widely used algorithms for min-max optimization. More precisely, we compare: Gradient Descent Ascent (GDA; Figure 4), Extra-Gradient (EG; Figure 5), Follow-the-Ridge (FtR; Figure 6), and STay-ON-the-Ridge (STON'R; Figure 7) in the following 2-D examples:

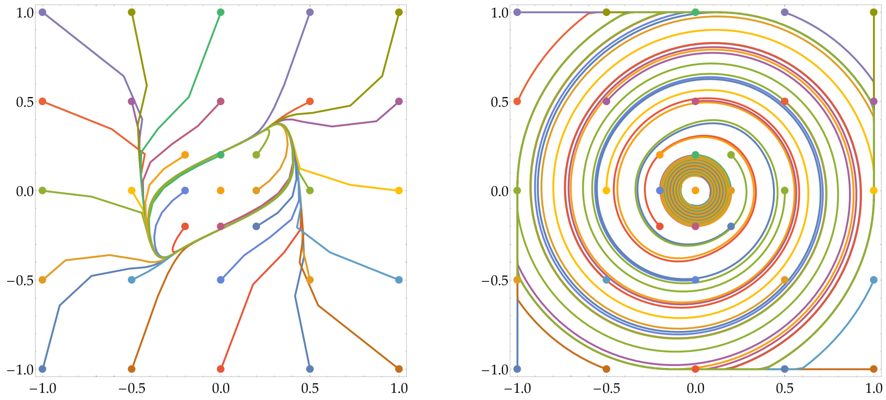
$$\min_{\theta\in[-1,1]} \max_{\omega\in[-1,1]} f_1(\theta, \omega) := (4\theta^2 - (\omega - 3\theta + \frac{\theta^3}{20})^2 - \frac{\omega^4}{10}) \exp(-\frac{\theta^2 + \omega^2}{100}), \text{ and}$$

$$\min_{\theta\in[-1,1]} \max_{\omega\in[-1,1]} f_2(\theta, \omega) := -\theta\omega - \frac{1}{20} \cdot \omega^2 + \frac{2}{20} \cdot S\left(\frac{\theta^2 + \omega^2}{2}\right) \cdot \omega^2$$

where S is the smooth-step function $S(\theta) = \begin{cases} 0, \theta \leq 0 \\ 3\theta^2 - 2\theta^3, \theta \in [0, 1] \\ 1, \theta \geq 1 \end{cases}$. Observe that in both cases it is easy to check that the only local min-max equilibrium is at $(0, 0)$.

We do not provide separate plots for Optimistic Gradient Descent Ascent (OGDA) because its behavior is almost identical with the behavior of EG in these examples and hence all our comments about EG transfer to OGDA as well. In all the following figures the different colors represent trajectories with different initialization. The initialization of every trajectory is represented by a dot and the line represent the path that the algorithm follows starting from the dot.

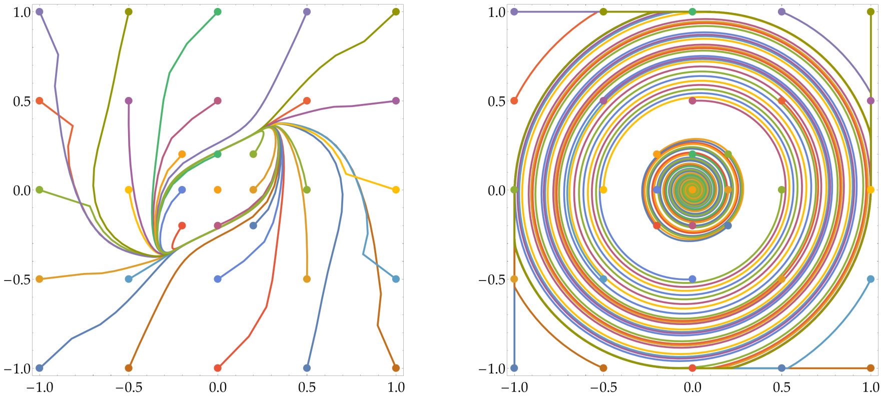
Observe that all the known methods either get trapped on a limit cycle, or they only converge when initialized very close to the solution. Our algorithm (Figure 7) is the only one that converges in both of these examples when initialized in $(-1, -1)$ which is far away from the solution.



(a) $f_1(\theta, \omega)$.

(b) $f_2(\theta, \omega)$.

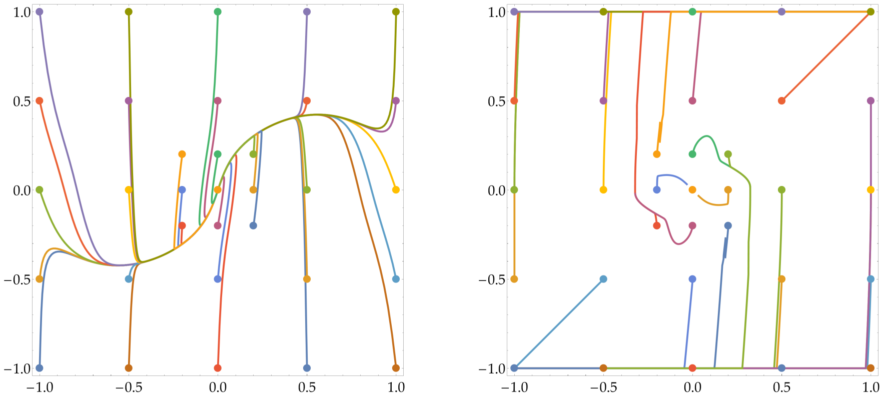
Figure 4: (Algorithm: GDA) **(a)** We observe that for any initial condition the algorithm converges to the same limit cycle. The only exception is when the algorithm is initialized exactly on $(0, 0)$ where the gradients are 0 and hence it does not move. So in this example, unless initialized on the equilibrium, the algorithm converges to a specific limit cycle. **(b)** In this example, if the algorithm is initialized far away from the equilibrium, which is $(0, 0)$, then it *diverges*, i.e., it moves towards the boundary. On the other hand, if the algorithm is initialized close enough to the equilibrium then it slowly converges to the equilibrium point with a very slow rate.



(a) $f_1(\theta, \omega)$.

(b) $f_2(\theta, \omega)$.

Figure 5: (Algorithm: EG/OGDA) **(a)** we observe that for every initial conditions the algorithm converges to the same limit cycle with the only exception of $(0, 0)$ as for GDA in Figure 4. **(b)** The behavior of the algorithm for $f_2(\theta, \omega)$ is again similar to the behavior of GDA as we can see in Figure 4 (b). There only two differences with GDA: (1) when initialized close to equilibrium, EG converges very fast, and (2) the region of attraction to the equilibrium is larger compared to GDA.



(a) $f_1(\theta, \omega)$.

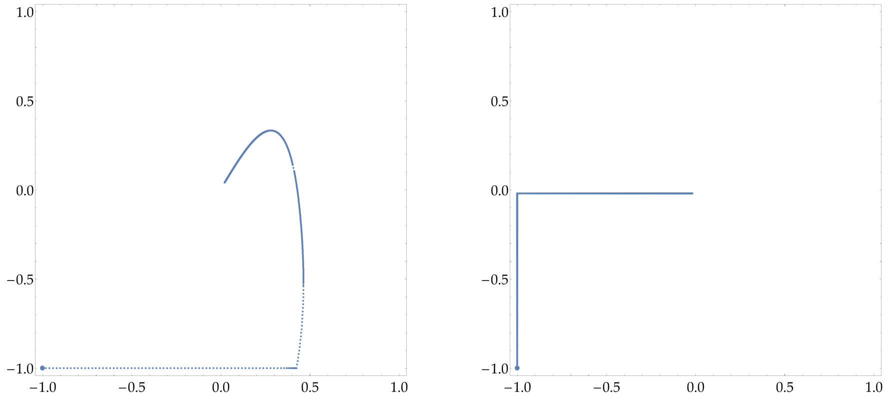
(b) $f_2(\theta, \omega)$.

Figure 6: (Algorithm: FtR) **(a)** We observe that for any initial condition the algorithm, in this example, converges to the equilibrium, in contrast with GDA or EG or OGDA. **(b)** In this example the behavior of the algorithm is very similar with GDA or EG or OGDA. If the algorithm is initialized far away from the equilibrium then it converges to either $(1, 1)$ or $(-1, -1)$ and none of them are equilibrium points. It is only when the algorithm is initialized next to the equilibrium that it converges to the equilibrium. Moreover, the algorithm needs to be initialized even closer than GDA to guarantee convergence. On the other hand, if the algorithm is initialized next to the equilibrium then it converges extremely fast, even faster than EG.

D PROOF OF LEMMA 1

Proof. $(\longleftarrow)$ Let Z denote the zero-satisfied coordinates ($V_i(x) = 0$), BS^+ the boundary satisfied coordinates with $x_i = 1$ (and thus $V_i(x) > 0$) and BS^- the boundary satisfied coordinates with $x_i = 0$ (and thus $V_i(x) < 0$). For any $y \in [0, 1]^n$, we have $\sum_{i=1}^{n} V_i(x)(x_i - y_i) \geq 0$, which can easily be seen by breaking up the sum into three sums corresponding to indices in Z, BS^+ and BS^-.

$(\longrightarrow)$ Let $x \in [0, 1]^n$ be a solution of the V, i.e. $V(x)^\top(x - y) \leq 0$ for all $y \in [0, 1]^n$. Consider an arbitrary $i \in [n]$ and a vector y such that $y_j = x_j$ for all $j \neq i$. If $x_i = 1$, take $y_i = 0$, and plug this into $V(x)(x - y) \leq 0$ to get $V_i(x) \geq 0$. If $x_i = 0$, take $y_i = 1$, and plug this into $V(x)^\top(x - y) \leq 0$ to get $V_i(x) \leq 0$. If $x_i \in (0, 1)$ consider first $y_i = x_i + \delta$ for some small $\delta > 0$



(a) $f_1(\theta, \omega)$.

(b) $f_2(\theta, \omega)$.

Figure 7: (Algorithm: STON'R) The STON'R algorithm is always initialized at $(-1, -1)$ independently of the objective function f. Hence, there is a good initialization for STON'R that is trivial to compute. This is contrast with the FtR algorithm that requires to be initialized close to the equilibrium. Such initialization might be as difficult to compute as finding the equilibrium itself. **(a)** we observe that the algorithm converges to the equilibrium almost directly and in particular it does not even need to spiral around the equilibrium. **(b)** The same for this example as well. The algorithm converges very fast and directly to the equilibrium although it is initialized far away from it. To the best of our knowledge, none of the known algorithms can achieve such a converge guarantee in this example.

and plug this into $V(x)^\top(x - y) \leq 0$ we get $V(x_i) \geq 0$. By repeating the same argument for $y_i = x_i - \delta$ we that get $V_i(x) \leq 0$. As a result, $V_i(x) = 0$. □

E PROOF OF THEOREM 1

In this section we present the proof of Theorem 1. The proof follows closely the sketch exhibited in Section 5.4 with some slight modifications on the definition of the nodes N of the directed graph G.

E.1 HELPFUL DEFINITIONS AND LEMMAS

We start with the definition of pivots that will play the role of nodes N.

Definition 4. *A point $x \in [0, 1]^n$ is called a pivot if and only if the following hold,*

- *If coordinate i is not satisfied then $V_i(x) > 0$.*
- *If ℓ is the minimum unsatisfied coordinate then $x_j = 0$ for all coordinates $j \geq \ell + 1$.*
- *If ℓ is the minimum unsatisfied coordinate then there exists at least one coordinate $j \in M \cup \{\ell\}$ with $x_j = 0$ or $x_j = 1$ where $M := \{j \leq \ell - 1 : \; V_j(x) = 0\}$.*

As in the proof sketch of Section 5.4, given a pivot x (that admits at least one unsatisfied variable) we argue that STON'R visits another pivot at some finite time. As depicted in Dynamics 2, the latter happens by following the continuous curve $\dot{z}(t) = D^i_S(z(t))$. Recall that in Dynamics 2 the pair (i, S) is updated in the previous steps of the algorithm (Steps $8, 13, 15$ and 17). In the next Definitions 5, 6 and 7, we provide an alternative way of "computing locally" the pair (i, S) by using only the knowledge of $x(t)$ at Step 3 of Dynamics 2.

Definition 5. *Consider the direction $D^i_S(x) := (d_1, \ldots, d_n)$ of Definition 2 for the set of zero-satisfied coordinates $S = \{j < i \text{ with } V_j(x) = 0\}$ (recall that $d_j = 0$ for all $j \notin S \cup \{i\}$). If, for all $k \in S$, one of the following holds: (a) $x_k \in (0, 1)$, or (b) $x_k = 0$ and $d_k \geq 0$, or (c) $x_k = 1$ and $d_k \leq 0$, then we define $D^i(x) := D^i_S(x)$. Otherwise, let $j \in S$ be the unique coordinate (uniqueness follows from Assumption 2) such that either $\{x_j = 0$ and $d_j < 0\}$ or $\{x_j = 1$ and $d_j > 0\}$, and we define $D^i(x) := D^i_{S \setminus \{j\}}(x)$. $D^i(x)$ is called the* ideal direction of movement *at point $x \in [0, 1]^n$ with respect to coordinate i.*

Definition 6. *Given a point $x \in [0,1]^n$ coordinate i is called frozen if and only if ($x_i = 0$ and $[D^i(x)]_i < 0$) or ($x_i = 1$ and $[D^i(x)]_i > 0$) where $D^i(x)$ is the ideal direction at x with respect to coordinate i (Definition 5).*

Definition 7. *Given a pivot $x \in [0,1]^n$ consider*

- $\ell := \min_{1 \leq j \leq n}\{$*coordinate j is not satisfied at x*$\}$.
- $i := \max_{j \leq \ell}\{$*coordinate j is not frozen at x*$\}$.
- $S \leftarrow$ *the set of coordinates such that $D^i(\cdot) = D^i_S(\cdot)$ (see Definition 5).*

The coordinate i is called the under examination coordinate, the pair (i,S) is called the admissible pair for pivot x.

Remark 2. *Computing the (i,S) admissible pair of the pivot $x(t)$ at Step 3 in Dynamics 2 is equivalent with Dynamics 2 at which (i,S) is updated at Steps 8, 13, 15 and 17.*

E.2 MAIN STEPS OF THE PROOF

To simplify notation we describe STON'R using the notion of pivots and admissible pairs (i,S) of Definition 4 and 7.

Dynamics 3 STay-ON-the-Ridge (STON'R)

```
1: Initially $x(0) \leftarrow (0,\ldots,0)$, $i \leftarrow 1$, $S \leftarrow \emptyset$, $t \leftarrow 0$.
2: while $x(t)$ is not a VI solution do
3:     At point $x(t)$ compute the admissible pair $(i,S)$ for pivot $x(t)$.
4:     Follows the continuous-time dynamics, $\dot{z}(\tau) = D^i_S(z(\tau))$, at $z(0) = x(t)$.
5:     while $z(\tau)$ is not an exit point as per Definition 3 do
6:         Execute $\dot{z}(\tau) = D^i_S(z(\tau))$ forward in time.
7:     end while
8:     Set $x(t+\tau) = z(\tau)$ for all $\tau \in [0,\tau_{\text{exit}}]$ (where $\tau_{\text{exit}}$ is earliest time $z(\tau)$ became an exit point).
9:     Set $t \leftarrow t + \tau_{\text{exit}}$.
10: end while
11: return $x(t)$
```

We are now ready to present the topological argument described in Section 5.4. As already mentioned the nodes N of the directed graph G will be the set of pivots while we say that there exists an edge (x,x') from pivot x to pivot x' in case setting $z(0) := x$ and following the direction $\dot{z}(t) = D^i_S(z(t))$ (where (i,S) is the admissible pair of x) leads to pivot x' once one of the "if" loops in Steps 9, 11, 13 and 15 is activated.

In Lemma 3 we establish the fact that the pivots which correspond to the number of nodes of directed graph G are finite.

Lemma 3. *There exists a finite number of pivots.*

In Definition 8 we formalize the notion of directed edge (x,x') in graph G which we additionally denote as $x' = \mathrm{Next}(x)$.

Definition 8. *Given a pivot $x \in [0,1]^n$ consider the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x$ where (i,S) is the admissible pair of x. We say that pivot x' is the next pivot of x, i.e. $x' = \mathrm{Next}(x)$ if and only if there exists $t^* > 0$ such that*

- $z(t^*) = x'$
- *$z(t)$ is not a pivot for all $t \in (0,t^*)$.*

In Lemma 4 we establish the fact that any pivot with at least one unsatisfied variable must necessarily admit outdegree equal to 1. The latter directly implies that any pivot with outdegree 0 must correspond to a solution since all coordinates are satisfied.

Lemma 4. *For any pivot $x \in [0,1]^n$ with at least one unsatisfied coordinate there exists a pivot x' such that $x' = \mathrm{Next}(x)$. Moreover let (i, S) be the admissible pair for pivot x, $z(t)$ be the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x$ and t' be the time at which $x' = z(t')$. Then for all $t \in [0, t']$,*

- *all coordinates $j \in S$ admit $V_j(z(t)) = 0$.*
- *all coordinates $j \leq i - 1$ are satisfied at $z(t)$.*
- *all coordinates $j \geq i + 1$ admit $z_j(t) = 0$.*
- *all coordinates j admit $z_j(t) \in [0, 1]$.*

Using Lemma 4 we additionally obtain Corollary 1 ensuring that the point $x(t)$ at Step 3 of Dynamics 3 is always a pivot and thus Dynamics 3 is well-defined.

Corollary 1. *Let $x(t)$ at Step 3 of Dynamics 3 be a pivot. Then the point $x(t + \tau_{exit})$ at Step 8 of Dynamics 3 is also a pivot. Moreover the point $(0, \dots, 0)$ is a pivot.*

In Lemma 5 we establish the fact that no pivot/node can admit in-degree more than 2. The latter implies if we start with a pivot with 0 in-degree we must essentially visit a pivot with out-degree 0 that consists a solution.

Lemma 5. *Any pivot $x \in [0,1]^n$ admits in-degree at most 1. In other words in case $x^* = \mathrm{Next}(x_1)$ and $x^* = \mathrm{Next}(x_2)$ for some pivots x_1, x_2 then $x_1 = x_2$.*

We conclude the proof by showing that $(0, \dots, 0)$ that is the initial pivot that Dynamics 3 visits admits 0 in-degree.

Lemma 6. *There is no pivot $x \in [0,1]^n$ such that $\mathrm{Next}(x) = (0, \dots, 0)$.*

F PROOF OF LEMMA 3

Lemma 7. *Let the functions $F_1(x), \dots, F_i(x)$ where $F_\ell : [0,1]^i \mapsto \mathbb{R}$ and the set $B := \{x \in [0,1]^i : \; F_\ell(x) = 0$ for all $\ell = 1, \dots, i\}$. In case $F_1, \dots, F_\ell$ satisfy the following assumptions*

- $\|\nabla F_\ell(x) - \nabla F_\ell(y)\|_2 \leq L \cdot \|x - y\|_2$
- *For all $x \in B$ the matrix*

$$J(x) := \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \dots & \frac{\partial F_1(x)}{\partial x_i} \\ \vdots & & \vdots \\ \frac{\partial F_i(x)}{\partial x_1} & \dots & \frac{\partial F_i(x)}{\partial x_i} \end{pmatrix}$$

  admits singular values that are at least σ_{min} and at most σ_{max}.

Then the set B is finite. More precisely, $|B| \leq 2^i / \mathrm{Vol}^i\left(\frac{2\sigma^2_{min}}{\sqrt{i} L \sigma^2_{max}}\right)$ where $\mathrm{Vol}^i(\rho)$ is the volume of the i-dimensional ball with radius ρ.

Lemma 3 follows by Lemma 7. By Definition 4 a pivot admits m coordinates on the boundary and $n - m$ coordinates that are zero-satisfied. By fixing a specific set of coordinates (of size m) to be on the boundary, the rest of the coordinates must satisfy $V_\ell(x) = 0$. Let x_m denote the $\{0,1\}$-assignment of the m-boundary coordinates and apply Lemma 8 with $F_\ell(\cdot) := V_\ell(\cdot, x_m)$. Then we get that there is a finite set of points $z \in [0,1]^{n-m}$ such that $V_\ell(z, x_m) = 0$ for each of the rest $n - m$ coordinates (by Lemma 7 the number of such z is at most $2^n / \mathrm{Vol}^n\left(\frac{2\sigma^2_{min}}{\sqrt{n} L \sigma^2_{max}}\right)$). Since they are only 2^n choices for the boundary coordinates, the overall number of pivots is at most $4^n / \mathrm{Vol}^n\left(\frac{2\sigma^2_{min}}{\sqrt{n} L \sigma^2_{max}}\right)$.

F.1 PROOF OF LEMMA 7

Let us assume the existence of $x, y \in B$ such that $\|x - y\|_2 \leq \rho$ and $x \neq y$. Notice that the $\nabla F_1(x), \ldots, \nabla F_i(x)$ are linearly independent and thus

$$x - y = \sum_{j=1}^{i} \mu_j \cdot \nabla F_j(x)$$

which implies that

$$\|\mu\|_2 \leq \frac{\rho}{\sigma_{\text{min}}} \tag{5}$$

By Taylor expansion of x and the fact that $\|\nabla F_\ell(x) - \nabla F_\ell(y)\| \leq L \cdot \|x - y\|_2$ we get,

$$\left| F_\ell(y) - F_\ell(x) - (\nabla F_\ell(x))^\top \cdot \sum_{j=1}^{i} \mu_j \cdot \nabla F_j(x) \right| \leq \frac{1}{2} L \cdot \|\sum_{\ell=1}^{i} \mu_j \cdot \nabla F_j(x)\|^2$$

which due to the fact that $F_\ell(y) = F_\ell(x) = 0$ implies,

$$\left[J^\top(x) \cdot J(x) \cdot \mu \right]_\ell \leq \frac{1}{2} L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|^2$$

and thus

$$\|\mu\|_2 \geq \frac{2\sigma_{\text{min}}}{\sqrt{i} L \sigma_{\text{max}}^2} \tag{6}$$

Combining Equation 5 and 6 we get $\rho \geq \frac{2\sigma_{\text{min}}^2}{\sqrt{i} L \sigma_{\text{max}}^2}$. To this end we know that in case $x, y \in B$ with $x \neq y$ then $\|x - y\|_2 \geq \frac{2\sigma_{\text{min}}^2}{\sqrt{i} L \sigma_{\text{max}}^2}$. Thus,

$$|B| \leq 2^i / \text{Vol}^i \left(\frac{2\sigma_{min}^2}{\sqrt{i} L \sigma_{max}^2} \right)$$

G PROOF OF LEMMA 4

Lemma 8. *Let a pivot $x \in [0,1]^n$ and (i, S) the admissible pair for x. Then the following hold,*

- *There exists a unique trajectory $z(t)$ with $\dot{z}(t) = D_S^i(z(t))$ and $z(0) = x$.*
- *$V_j(z(t)) = 0$ for all coordinates $j \in S$.*
- *There exists $t^* > 0$ such that for all $t \in [0, t^*]$ all coordinates $j \leq i - 1$ are satisfied at $z(t)$ and $z_j(t) \in [0,1]$ for all coordinates j.*

Lemma 9. *Let a set of coordinates S, a coordinate i and a point $x \in [0,1]$ such that $x_j \in (0,1)$ for all $j \in S \cup \{i\}$. Consider the trajectory $\dot{\gamma}(t) = D_S^i(\gamma(t))$ with $\gamma(0) = x$. Then there exists $t^* \in (0, C]$ such that*

$$\gamma_j(t^*) = 0 \text{ or } 1 \quad \text{for some } j \in S \cup \{i\}$$

where C is constant depending on the parameters $\sigma_{min}, \sigma_{max}$ and L.

Lemma 10. *Let a pivot $x \in [0,1]^n$ with at least one unsatisfied coordinate. Let (i, S) the admissible pair of pivot x (Definition 7) and ℓ the minimum unsatisfied coordinate at x. Then the following hold,*

- *The under examination variable admits $i \geq 1$.*
- *$x_j = 0$ for all coordinates $j \geq i + 1$.*

Additionally one of the following holds,

- *$i = \ell$ and $V_\ell(x) > 0$*
- *$x_i = 1$ and $V_i(x) > 0$*
- *$V_i(x) = 0$ and $D_S^i(x)^\top \nabla V_i(x) > 0$*

G.1 PROOF OF LEMMA 4

Given the pivot $x \in [0,1]^n$ with at least one unsatisfied variable and let (i,S) denote its admissible pair. By Lemma 10 we know that the under examination variable i admits $i \geq 1$. Now consider the the trajectory $\dot{z}(t) = D_S^i(z(t))$ with $z(0) = x$. Due to the fact that $i \geq 1$ and by Assumption 3 we known that for all $t \in (0,\delta)$ where $\delta > 0$ is sufficiently small, the following hold

- $z_j(t) \in (0,1)$ for all $j \in S \cup \{i\}$
- $z_j(t) = 0$ or $z_j(t) = 1$ for all coordinates $j \notin S \cup \{i\}$.

By Lemma 9 there exists $t^* > 0$ such that $z_j(t^*) = 0$ or 1 for some coordinate $j \in S \cup \{i\}$ and $z_j(t) \in [0,1])$ for all coordinates j and $t \in [0,t^*]$.

We first show that if there exists a coordinate $j \leq i-1$ such that j is not satisfied at $z(t^*)$ then there exists $\hat{t} < t^*$ such that $z(\hat{t})$ is a pivot.

Let ℓ denote the unsatisfied coordinate at $z(t^*)$. Notice that by Lemma 8 all coordinates $j \in S$ admit $V_j(z(t^*)) = 0$ and thus $\ell \notin S$. The latter implies that ($x_\ell = 0$ and $V_\ell(x) \leq 0$) or ($x_\ell = 1$ and $V_\ell(x) \geq 0$) and since coordinate ℓ stands still in the trajectory $\dot{z}(t) = D_S^i(z(t))$, $z_\ell(\hat{t}) = x_\ell$ there are two mutually exclusives cases:

- $x_\ell = z_\ell(\hat{t}) = 0$, $V_\ell(x) \leq 0$ and $V_\ell(z(\hat{t})) > 0$
- $x_\ell = z_\ell(\hat{t}) = 1$, $V_\ell(x) \geq 0$ and $V_\ell(z(\hat{t})) < 0$

Then by Lemma 12 we additionally get that for sufficiently small $\delta > 0$,

- If $x_\ell = 0$ then $V_\ell(z(t)) < 0$ for $t \in (0,\delta)$
- If $x_\ell = 1$ then $V_\ell(z(t)) > 0$ for $t \in (0,\delta)$

As a result, in any case there exists $t_\ell \in (0,t^*)$ such that $V_\ell(z(t_\ell)) = 0$, coordinate ℓ lies on the boundary at $z(t_\ell) = 0$ and coordinate ℓ is satisfied at $z(t)$ for all $t \in [0,t_\ell]$.

Now consider the set of coordinates $A := \{j \leq i-1 :$ coordinate j is not satisfied at $z(t^*)\}$ and let $\hat{t} := \min_{\ell \in A} \hat{t}_\ell$. Then all coordinates $j \leq i-1$ are satisfied at $z(\hat{t})$ while there exists a coordinate $\hat{\ell} \in A$ such that $V_{\hat{\ell}}(z(\hat{t})) = 0$ with coordinate $\hat{\ell}$ being on the boundary at $z(\hat{t})$. Up next we argue that $z(\hat{t})$ is a pivot.

Consider the set of coordinates $M := \{j \leq i-1 :\ V_j(z(\hat{t})) = 0\}$. Since $\hat{\ell} \in M$ and $\hat{\ell}$ lies on the boundary at $z(\hat{t})$ the third item of Definition 4 is satisfied. Since x is a pivot, Lemma 10 implies all coordinates $j \geq i+1$ admit $x_j = 0$. Since $[D_S^i(\cdot)]_j = 0$ for all $j \geq i+1$ the latter implies that $z_j(\hat{t}) = x_j = 0$. Due to the fact that all coordinates $j \leq i-1$ are satisfied at $z(\hat{t})$ we get that the second item of Definition 4 is satisfied. Finally notice that by Lemma 10, $V_i(z(t)) > 0$ for all $t \in (0,\delta)$ for sufficiently small δ. In case $V_i(z(\hat{t})) < 0$ then there exist $t' < \hat{t}$ such that $V(z(t')) = 0$ implying that $z(t')$ is a pivot. As a result, without loss of generality we assume that $V_i(z(\hat{t})) > 0$ which implies that the first item of Definition 4 is satisfied.

As a result, without loss of generality we assume that all coordinates $j \leq i-1$ are satisfied for all $z(t)$ in $[0,t^*]$. Up next we show that in this case either the point $x^* := z(t^*)$ is a pivot or $z(\hat{t})$ is pivot for some $\hat{t} \in (0,t^*)$.

Notice that by Lemma 10 all coordinates $j \geq i+1$ admit $x_j = 0$. Since $[D_S^i(\cdot)]_j = 0$ for all $j \geq i+1$ the latter implies that $x_j^* = x_j = 0$. As a result, $x_j^* = 0$ for all $j \geq i+1$. Due to our assumption that all coordinates $j \leq i-1$ are satisfied at x^*, the minimum unsatisfied coordinate at x^* is greater than i and thus the second item of Definition 4 is satisfied. Moreover due to the fact that $x_j^* = 0$ or 1 for some $j \in S \cup \{i\}$ and $V_j(x^*) = 0$ for all $j \in S$, the third item of Definition 4 is satisfied.

Up next we argue that in case coordinate i is not satisfied at x^* then $V_i(x^*) > 0$. Let us assume that coordinate i is not satisfied at x^* and $V_i(x^*) < 0$. Let ℓ denote the minimum unsatisfied variable at x then Lemma 10 provides us with the following mutually exclusive cases:

- $i = \ell$ then $V_\ell(x) > 0$: Since $V_\ell(z(0)) = V_\ell(z) > 0$ and $V_\ell(z(t^*)) = V_\ell(x^*) < 0$ there exists $\hat{t} \in (0, t^*)$ such that $V_\ell(z(\hat{t})) = 0$. Notice that $z(\hat{t})$ satisfies all the three items of Definition 4.
- $x_i = 1$ with $V^i(x) > 0$: Same as above.
- $V_i(x) = 0$ and $D^i_S(x)^\top \cdot V^i(x) > 0$: Notice that $V_i(z(t)) > 0$ for all $t \in (0, \delta)$ once δ is selected sufficiently small. By repeating the same argument as above we conclude that there is $\hat{t}$ such that $z(\hat{t})$ is a pivot.

G.2 PROOF OF LEMMA 8

Let the set of coordinates S, we first establish in Lemma 11 that $D^i_S(x)$ is M-Lipschitz. The proof of Lemma 11 is presented at Section G.4

Lemma 11. *Let $x \in [0,1]^n$ and a set of coordinates S such that $V_j(x) = 0$ for all $j \in S$. Then for any coordinate $i \notin S$ and for any $y \in \mathbb{R}^n$ such that $\|x - y\|_2 \leq \sigma_{min}/2L\sqrt{n}$,*

$$\|D^i_S(x) - D^i_S(y)\|_2 \leq M \cdot \|x - y\|_2$$

for $M := \Theta\left(\frac{\sigma_{max}}{\sigma^2_{min}} \cdot \sqrt{n} \cdot L\right)$.

To simplify notation let $Z^i(x) := \{\ell < i \text{ such that } V_\ell(x) = 0\}$ and $F^i(x) := \{\ell < i \text{ such that coordinate } \ell \text{ is boundary satisfied at } x\}$. Since x is a pivot, all coordinates $\ell \leq i - 1$ are satisfied and thus each coordinate $\ell \leq i-1$ either belongs to $Z^i(x)$ or to $F^i(x)$.

Let us first consider the case where one of the following holds for all coordinates $\ell \in Z^i(x)$.

- $x_\ell \in (0,1)$
- $x_\ell = 0$ and $[D^i_{Z_i(x)}]_\ell \geq 0$
- $x_\ell = 1$ and $[D^i_{Z_i(x)}]_\ell \leq 0$

Notice that in this case Definition 5 and 7 imply $S = Z^i(x)$. Now consider the set $B := \{y \in \mathbb{R}^n \text{ such that } \|x - y\|_2 \leq \sigma_{\min}/2L\}$. Then combining Lemma 11 with the Picard–Lindelöf theorem we get that there exists a unique $z(t)$ such that

1. $\dot{z}(t) = D^i_S\left(z(t)\right)$
2. $z(0) = x$

By taking $\delta > 0$ sufficiently small we get that for all $t \in [0, \delta]$ the following hold,

- $V_\ell\left(z(t)\right) = 0$ for all $\ell \in S$
- $z_\ell(t) \in (0,1)$ for all $\ell \in S$.
- coordinate ℓ is boundary satisfied at $z(t)$ for all $\ell \in \{1, \ldots, i-1\}/S$.
- $z_j(t) \in [0,1]^n$ for all coordinates j.

Now consider the case in which there exists a coordinate $j \in Z^i(x)$ such that ($x_j = 0$ and $[D^i_{Z_i(x)}]_j < 0$) or ($x_j = 1$ and $[D^i_{Z_i(x)}]_j > 0$). By Assumption 2 we know that such a coordinate must be unique. In this case by Definition 5 we get $D^i(x) = D^i_{Z^i(x)/\{j\}}(x)$ and thus by Definition 7, $S = Z^i(x)/\{j\}$.

Lemma 12 establish the fact that in this case following the direction $D^i_S(x)$ consists the variable j boundary satisfied. The proof of Lemma 12 is presented in Section G.3.

Lemma 12. *For any $x \in [0,1]^n$ if there exists coordinate j with*

- *$x_j = 0$ and $[D^i_{Z^i(x)}(x)]_j < 0$ then $\left(D^i_{Z^i(x)/\{j\}}(x)\right)^\top \cdot \nabla V_j(x) < 0$.*
- *$x_j = 1$ and $[D^i_{Z^i(x)}(x)]_j > 0$ then $\left(D^i_{Z^i(x)/\{j\}}(x)\right)^\top \cdot \nabla V_j(x) > 0$.*

By the exact same arguments as above, we get that there exists a unique trajectory $z(t)$ such that $\dot{z}(t) = D^i_S\left(z(t)\right)$ and $x(0) = x$ and by taking $\delta > 0$ sufficiently small we get,

- $V_\ell\left(z(t)\right) = 0$ for all $\ell \in S$
- $z_\ell(t) \in (0, 1)$ for all $\ell \in S$
- coordinate ℓ is boundary satisfied at $z(t)$ for all $\ell \in F^i(x)$
- $z_j(t) \in [0, 1]^n$ for all coordinates j.

In order to complete the proof of Lemma 8 we need to argue that the coordinate j is satisfied for all $z(t)$ with $t \in [0, \delta']$. Without loss of generality consider $x_j = 0$ (the case $x_j = 1$ follows symmetrically). Recall that $V_j(x) = 0$ and by Lemma 12 we get that $\left(D^i_S(x)\right)^\top \cdot \nabla V_j(x) < 0$. Thus by selecting $\delta' < \delta$ sufficiently small we get

$$V_j\left(z(t)\right) < 0 \quad \text{and} \quad z_j(t) = 0$$

for all $t \in (0, \delta']$.

G.3 PROOF OF LEMMA 12

To simplify notation let $Z^i(x) = \{1, \ldots, i-1\}$, $D^i_{Z^i(x)}(x) = (d_1, \ldots, d_j, \ldots, d_i)$ and $D^i_{Z^i(x)/\{j\}}(x) = (\hat{d}_1, \ldots, \hat{d}_{j-1}, \hat{d}_{j-1}, \ldots, \hat{d}_i)$. Moreover let assume that $x_j = 0$ and i is even. The cases $x_j = 0$ and i is odd, $x_j = 1$ and i is even, $x_j = 1$ and i is odd follow symmetrically.

We will prove that

$$\left(\hat{d}_1, \ldots, \hat{d}_{j-1}, \hat{d}_{j+1}, \ldots, \hat{d}_i\right)^\top \cdot \left(\frac{\partial V_j(x)}{\partial x_1}, \ldots, \frac{\partial V_j(x)}{\partial x_{j-1}}, \frac{\partial V_j(x)}{\partial x_{j+1}}, \ldots, \frac{\partial V_j(x)}{\partial x_{i-1}}\right) < 0$$

Since i is even we get by Definition 2,

$$\begin{vmatrix}
\frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \hat{d}_1 \\
\vdots & & \vdots & & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_{j-1}} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & \hat{d}_{j-1} \\
\frac{\partial V_1(x)}{\partial x_{j+1}} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & \hat{d}_{j+1} \\
\vdots & & \vdots & \vdots & \vdots & & \vdots \\
\frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \hat{d}_i
\end{vmatrix} > 0 \tag{7}$$

and that

$$\begin{vmatrix}
\frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_j(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_{j-1}} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_j(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & d_{j-1} \\
\frac{\partial V_1(x)}{\partial x_j} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_j} & \frac{\partial V_j(x)}{\partial x_j} & \frac{\partial V_{j+1}(x)}{\partial x_j} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_j} & d_j \\
\frac{\partial V_1(x)}{\partial x_{j+1}} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_j(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & d_{j+1} \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_j(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_i} & d_i
\end{vmatrix} < 0 \tag{8}$$

Combining the fact that $\left(\frac{\partial V_\ell(x)}{\partial x_1}, \dots, \frac{\partial V_\ell(x)}{\partial x_i}\right)^\top \cdot (d_1, \dots, d_i) = 0$ (see Definition 2) with $d_j < 0$ (we have assumed that $x_j = 0$) we get by Equation 8,

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_j(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_j(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & d_{j-1} \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & d_1^2 + \dots + d_i^2 \\ \frac{\partial V_1(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_j(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & d_{j+1} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_j(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & d_i \end{vmatrix} > 0 \tag{9}$$

which implies that

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \frac{\partial V_j(x)}{\partial x_1} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & \frac{\partial V_j(x)}{\partial x_{j-1}} \\ \frac{\partial V_1(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & \frac{\partial V_j(x)}{\partial x_{j+1}} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \frac{\partial V_j(x)}{\partial x_i} \end{vmatrix} < 0 \tag{10}$$

Multiplying with Equation 7 we get,

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_1(x)}{\partial x_i} \\ \vdots & \vdots & \vdots \\ \frac{\partial V_{j-1}(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} \\ \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{j+1}(x)}{\partial x_i} \\ \vdots & \vdots & \vdots \\ \frac{\partial V_{i-1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} \\ \hat{d}_1 & \dots & \hat{d}_i \end{vmatrix} \cdot \begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \frac{\partial V_j(x)}{\partial x_1} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & \frac{\partial V_j(x)}{\partial x_{j-1}} \\ \frac{\partial V_1(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & \frac{\partial V_j(x)}{\partial x_{j+1}} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \frac{\partial V_j(x)}{\partial x_i} \end{vmatrix} < 0$$

Now using the fact that $\left(\frac{\partial V_\ell(x)}{\partial x_1}, \dots, \frac{\partial V_\ell(x)}{\partial x_{j-1}}, \frac{\partial V_\ell(x)}{\partial x_{j+1}}, \dots, \frac{\partial V_\ell(x)}{\partial x_i}\right)^\top \cdot (\hat{d}_1, \dots, \hat{d}_{j-1}, \hat{d}_{j+1}, \dots, \hat{d}_i) = 0$ (see Definition 2) implies that

$$\begin{vmatrix} \Phi_1^\top(x) \cdot \Phi_1(x) & \Phi_1^\top(x) \cdot \Phi_2(x) & \dots & \Phi_1^\top(x) \cdot \Phi_{i-1}(x) & A_1 \\ \Phi_2^\top(x) \cdot \Phi_1(x) & \Phi_2^T(x) \cdot \Phi_2(x) & \dots & \Phi_2^T(x) \cdot \Phi_{i-1}(x) & A_2 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ \Phi_{i-1}^\top(x) \cdot \Phi_1(x) & \Phi_{i-1}^\top(x) \cdot \Phi_2(x) & \dots & \Phi_{i-1}^\top(x) \cdot \Phi_{i-1}(x) & A_{i-1} \\ 0 & 0 & \dots & 0 & (\hat{d}_1, \dots, \hat{d}_i)^\top \cdot \left(\frac{\partial V_j(x)}{\partial x_1}, \dots, \frac{\partial V_j(x)}{\partial x_i}\right) \end{vmatrix} < 0$$

where $\Phi_\ell = \left(\frac{\partial V_\ell(x)}{\partial x_1}, \dots, \frac{\partial V_\ell(x)}{\partial x_{j-1}}, \frac{\partial V_\ell(x)}{\partial x_{j+1}}, \dots, \frac{\partial V_\ell(x)}{\partial x_i}\right)$. The latter implies Claim 12.

G.4 PROOF OF LEMMA 11

To simplify notation let $S := \{1, \dots, i-1\}$ and for $x \in [0,1]^n$ consider the matrix $A(x)$ and the vector $b(x)$

$$A(x) := \begin{pmatrix} \frac{\partial V_1(x)}{\partial x_1} & \frac{\partial V_2(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \\ \frac{\partial V_1(x)}{\partial x_2} & \frac{\partial V_2(x)}{\partial x_2} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_2} & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_{i-1}} & \frac{\partial V_2(x)}{\partial x_{i-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{i-1}} & \end{pmatrix} \quad \text{and} \quad b(x) := \begin{pmatrix} \frac{\partial V_1(x)}{\partial x_i} \\ \frac{\partial V_2(x)}{\partial x_i} \\ \vdots \\ \frac{\partial V_{i-1}(x)}{\partial x_i} \end{pmatrix}$$

Notice that since $V_j(x) = 0$ for all $j = 1, \dots, i-1$, Assumption 1 ensures that the matrix $A(x)$ admits singular value greater than $\sigma_{\min}$ and thus $A(x)$ is invertible. Moreover due to the fact that for all $x, y \in [0,1]^n$

$$\|\nabla V_j(x) - \nabla V_j(y)\|_2 \leq L \cdot \|x - y\|_2$$

we get that

$$\|A(x) - A(y)\|_2 \leq \sqrt{n}L \cdot \|x - y\|_2 \text{ and } \|b(x) - b(y)\|_2 \leq L \cdot \|x - y\|_2.$$

To simplify notation $C_A := \sqrt{n}L$ and $C_b := L$. Since $\|x - y\|_2 \leq \frac{\sigma_{\min}}{2\sqrt{n}L}$ we get that $A(y)$ admits singular value greater than $\sigma_{\min}/2$ and thus $A(y)$ is invertible.

Notice that the direction $D_S^i(x)$ of Definition 2 is either

$$\left(\frac{A^{-1}(x) \cdot b(x)}{\sqrt{1 + \|A^{-1}(x) \cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1 + \|A^{-1}(x) \cdot b(x)\|_2^2}}\right) \text{ or } \left(-\frac{A^{-1}(x) \cdot b(x)}{\sqrt{1 + \|A^{-1}(x) \cdot b(x)\|_2^2}}, -\frac{1}{\sqrt{1 + \|A^{-1}(x) \cdot b(x)\|_2^2}}\right)$$

depending on the sign of the determinant. We show that for an appropriately selected M,

$$\|\left(\frac{A^{-1}(x)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(y)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(y)\|_2^2}}\right)\|_2$$
$$\leq M \cdot \|x - y\|_2$$

In order to prove the above, we use a standard lemma in sensitivity analysis of linear systems.

Lemma 13. [2] *Let the invertible square matrices A, B such that $F := \|(A - B) \cdot A^{-1}\|_2 < 1$. Then,*

$$\frac{\|A^{-1}b - B^{-1}b\|_2}{\|A^{-1}b\|_2} \leq \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \cdot \frac{\|F\|_2}{1 - \|F\|_2}$$

We prove the following 4 inequalities,

- $\|\left(\frac{A^{-1}(x)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right)\|_2 \leq M_1 \cdot \|x - y\|_2$
- $\|\left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right)\|_2 \leq M_2 \cdot \|x - y\|_2$
- $\|\left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right)\|_2 \leq M_3 \cdot \|x - y\|_2$
- $\|\left(\frac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(y)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(y)\|_2^2}}\right)\|_2^2 \leq M_4 \cdot \|x - y\|_2$

and then Lemma 2 follows for $M := M_1 + M_2 + M_3 + M_4$.

Let the matrix $F := (A(x) - A(y)) \cdot A^{-1}(x)$ then the fact that $\|x - y\|_2 \leq \frac{\sigma_{\min}}{2C_A}$ implies,

$$\|F\|_2 = \|(A(x) - A(y)) \cdot A^{-1}(x)\|_2 \leq \frac{C_A}{\sigma_{\min}} \cdot \|x - y\|_2 \leq \frac{1}{2} \tag{11}$$

---

[2] https://www.colorado.edu/amath/sites/default/files/attached-files/linearsystems_0.pdf

For the first case we get,

$$
\begin{aligned}
&\|\left(\tfrac{A^{-1}(x)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\tfrac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right)\|_2^2 \\
&= \tfrac{\|A^{-1}(x)\cdot b(x) - A^{-1}(y)\cdot b(x)\|_2^2}{1+\|A^{-1}(x)\cdot b(x)\|_2^2} \\
&\leq \tfrac{\|A^{-1}(x)\cdot b(x) - A^{-1}(y)\cdot b(x)\|_2^2}{\|A^{-1}(x)\cdot b(x)\|_2^2} \\
&\leq \left(\tfrac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \cdot \tfrac{\|F\|_2}{1-\|F\|_2}\right)^2 \qquad \text{by Lemma 13} \\
&\leq \left(\tfrac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \cdot 2 \cdot \|F\|_2\right)^2 \qquad \text{by Equation 11} \\
&\leq \tfrac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} \cdot 4 \cdot \|(A(x) - A(y)) \cdot A^{-1}(x)\|_2^2 \leq 4\tfrac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^4} \cdot C_A^2 \cdot \|x-y\|_2^2
\end{aligned}
$$

Thus $M_1 := 2C_A \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}^2}$

For the second case

$$
\begin{aligned}
&\|\left(\tfrac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\tfrac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right)\|_2^2 \\
&= \left(\|A^{-1}(y)\cdot b(x)\|^2 + 1\right) \tfrac{\left(\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2} - \sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}\right)^2}{(1+\|A^{-1}(y)\cdot b(x)\|_2^2)\cdot(1+\|A^{-1}(x)\cdot b(x)\|_2^2)} \\
&\leq \left(\|A^{-1}(y)\cdot b(x)\|^2 + 1\right) \tfrac{\left(\|A^{-1}(x)\cdot b(x)\|_2 - \|A^{-1}(y)\cdot b(x)\|\right)^2}{(1+\|A^{-1}(y)\cdot b(x)\|_2^2)\cdot(1+\|A^{-1}(x)\cdot b(x)\|_2^2)} \qquad \text{since } \sqrt{1+b} - \sqrt{1+a} \leq \sqrt{b} - \sqrt{a} \\
&\leq \tfrac{\left(\|A^{-1}(x)\cdot b(x)\|_2 - \|A^{-1}(y)\cdot b(x)\|\right)^2}{\|A^{-1}(x)\cdot b(x)\|_2^2} \\
&\leq \tfrac{\|A^{-1}(x)\cdot b(x) - A^{-1}(y)\cdot b(x)\|_2^2}{\|A^{-1}(x)\cdot b(x)\|_2^2}
\end{aligned}
$$

Applying the exact same arguments as before, we get $M_2 := 2C_A \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}^2}$.

For the third case,

$$
\begin{aligned}
&\|\left(\tfrac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right) - \left(\tfrac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right)\|_2^2 \\
&= \tfrac{\|A^{-1}(y)\cdot b(y) - A^{-1}(y)\cdot b(x)\|_2^2}{1+\|A^{-1}(y)\cdot b(x)\|_2^2} \leq \tfrac{C_b^2}{\sigma_{\text{min}}^2} \cdot \|x-y\|_2^2
\end{aligned}
$$

For the forth case,

$$
\begin{aligned}
&\|\left(\tfrac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right) - \left(\tfrac{A^{-1}(y)\cdot b(y)}{\sqrt{1+\|A^{-1}(y)\cdot b(y)\|_2^2}}, \tfrac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(y)\|_2^2}}\right)\|_2^2 \\
&\leq \left(\|A^{-1}(y)\cdot b(y)\|_2^2 + 1\right) \cdot \tfrac{\|A^{-1}(y)\cdot b(y) - A^{-1}(y)\cdot b(x)\|^2}{(1+\|A^{-1}(y)\cdot b(y)\|^2)\cdot(1+\|A^{-1}(y)\cdot b(x)\|^2)} \\
&\leq \|A^{-1}(y)\cdot b(y) - A^{-1}(y)\cdot b(x)\| \leq \tfrac{C_b^2}{\sigma_{\text{min}}^2} \cdot \|x-y\|_2^2
\end{aligned}
$$

As a result, we overall get that $M := \Theta\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}^2} \cdot L \cdot \sqrt{n}\right)$.

G.5 PROOF OF LEMMA 9

To simplify notation let $S := (1, \ldots, i-1)$ and let $D_S^i(x)$ be denoted as $D(x)$. The existence and uniqueness of trajectory $\gamma(t)$ follows by the Picard–Lindelöf theorem and the fact that $D(x)$ is M-Lipschitz continuous (see the proof Lemma 8 and Lemma 11).

We also denote as $\Phi_\ell(x)$ the gradient of $V_\ell(x)$ with respect to the coordinates $\{1, \ldots, i\}$, i.e. $\Phi(x) := \left(\frac{\partial V_\ell(x)}{\partial x_1}, \ldots, \frac{\partial V_\ell(x)}{\partial x_i}\right)$. To simplify things we repeat the definition of $D(x)$ of Definition 2 with respect to the notation of this section.

Definition 9. *Given $x \in [0,1]^i$ the direction $D(x)$ is defined as follows,*

- $\nabla V_j(x)^\top \cdot (d_1, \dots, d_{i-1}, d_i) = 0$ *for all* $j = 1, \dots, i$.
- *the sign of* $\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \frac{\partial V_2(x)}{\partial x_1} & \cdots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\ \frac{\partial V_1(x)}{\partial x_2} & \frac{\partial V_2(x)}{\partial x_2} & \cdots & \frac{\partial V_{i-1}(x)}{\partial x_2} & d_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \frac{\partial V_2(x)}{\partial x_i} & \cdots & \frac{\partial V_{i-1}(x)}{\partial x_i} & d_i \end{vmatrix}$ *equals sign* $\left((-1)^{i-1}\right)$.
- $d_1^2 + \cdots + d_{i-1}^2 + d_i^2 = 1$.

Assumption 1 ensures that at any point $x \in [0,1]^i$ the matrix

$$\Phi(x) := \begin{pmatrix} \Phi_1(x) \\ \Phi_2(x) \\ \vdots \\ \Phi_{i-1}(x) \end{pmatrix} := \begin{pmatrix} \nabla V_1(x) \\ \nabla V_2(x) \\ \vdots \\ \nabla V_{i-1}(x) \end{pmatrix} \tag{12}$$

admits singular values greater than σ_{min} and smaller than σ_{max}.

Corollary 2. *For all $x, y \in [0,1]^i$ with $V_\ell(x) = V_\ell(y) = 0$ for $\ell \in \{1, \dots, i-1\}$,*

$$\|D(x) - D(y)\|_2 \leq M \cdot \|x - y\|_2$$

for $M := \Theta(\frac{\sigma_{max}}{\sigma_{min}^2} \cdot \sqrt{n} \cdot L)$.

Corollary 2 follows directly by Lemma 11. Up next we show that there exist a finite time $t^* > 0$ at which $\gamma(t)$ hits the boundary $[0,1]^i$.

Claim 1. *For each t_0, there exists $t \leq 1/M$ such that $\|\gamma(t + t_0) - \gamma(t_0)\|_2 \geq \frac{1}{4M}$.*

Proof. To simplify notation let $t_0 := 0$. and let us assume that $\|\gamma(t) - \gamma(0)\|_2 \leq \frac{1}{4M}$ for all $t \in [0, 1/M]$. The latter implies that for all $t_1, t_2 \in [0, 1/M]$,

$$\|\gamma(t_1) - \gamma(t_2)\|_2 \leq \frac{1}{2M}$$

which implies that for all $t_1, t_2 \in [0, 1/M]$

$$\|D(\gamma(t_1)) - D(\gamma(t_2))\|_2 \leq \frac{1}{2}.$$

Using the fact that $\|D(\gamma(t_1))\|_2 = \|D(\gamma(t_2))\|_2 = 1$ we get that,

$$D^\top(\gamma(t_1)) \cdot D(\gamma(t_2)) \geq 1/2$$

As a result,

$$\begin{aligned} \|\gamma(1/M) - \gamma(0)\|_2^2 &= \|\int_0^{1/M} D(\gamma(s))\, \partial s\|^2 \\ &= \int_0^{1/M} \int_0^{1/M} D^\top(\gamma(s)) \cdot D(\gamma(s'))\, \partial s\, \partial s' \geq \frac{1}{2M^2} \end{aligned}$$

and thus $\|\gamma(1/M) - \gamma(0)\|_2 \geq \frac{1}{\sqrt{2}M}$ which is a contradiction. □

Claim 2. *For any t_0, there exist $0 \leq t_1, t_2 \leq \frac{1}{M}$ such that*

1. $\|\gamma(t_0 + t_1) - \gamma(t_0)\|_2 \geq \frac{1}{4M}$.

2. $\|\gamma(t_0 - t_2) - \gamma(t_0)\|_2 \geq \frac{1}{4M}$

Proof. Symmetrically as Claim 1. $\square$

Lemma 14. *Let $\delta \leq 1/4$ and $\gamma \in [0,1]^n$ such that $\|\gamma(t_0) - \gamma\| \leq \frac{\delta}{2M}$. Then there exists $t^* \in [-1/M, 1/M]$ such that*

- $\|\gamma(t^* + t_0) - \gamma(t_0)\|_2 \leq \frac{\delta}{M}$.
- $D^\top(\gamma(t^* + t_0)) \cdot (\gamma(t^* + t_0) - \gamma) = 0$.

Proof. By Claim 2 there exists $0 \leq t_1 \leq 1/M$ such that $\|\gamma(t_1 + t_0) - \gamma(t_0)\| \geq \frac{1}{4M}$. Let $t' = \inf_{0 \leq t \leq 1/M}\{\|\gamma(t + t_0) - \gamma(t_0)\|_2 \geq \frac{\delta}{M}\}$. By the triangle inequality, we get that $\|\gamma(t' + t_0) - \gamma\|_2 \geq \frac{\delta}{2M}$ and thus there exists $\hat{t}_1 \in [0, t']$ such that

- $\|\gamma(\hat{t}_1 + t_0) - \gamma\|_2 = \frac{\delta}{2M}$
- $\|\gamma(t + t_0) - \gamma\|_2 < \frac{\delta}{2M}$ for $t \leq \hat{t}_1$.

The latter implies that

- $\|\gamma(t + t_0) - \gamma(t_0)\|_2 \leq \frac{\delta}{M}$ for all $0 \leq t \leq \hat{t}_1$.
- $D^\top\left(\gamma(t_0 + \hat{t}_1)\right) \cdot \left(\gamma(t_0 + \hat{t}_1) - \gamma\right) \geq 0$.

Symmetrically we can prove that there exists $\hat{t_2}$ such that

- $\|\gamma(t_0 - t) - \gamma(t_0)\|_2 \leq \frac{\delta}{M}$ for all $0 \leq t \leq \hat{t}_2$.
- $D^\top\left(\gamma(t_0 - \hat{t}_2)\right) \cdot (\gamma(t_0 - \hat{t}_2) - \gamma) \leq 0$.

The proof follows by continuity of $g(t) := D^\top\left(\gamma(t_0 + t)\right) \cdot \left(\gamma(t_0 + t) - \gamma\right)$ for $t \in [-\hat{t}_2, \hat{t}_1]$. $\square$

Up next we present the main lemma of the section.

Lemma 15. *Let $\rho = \Theta\left(\frac{\sigma_{min}^3}{\sqrt{n}\cdot\sigma_{max}^2\cdot L}\right)$ and a point $p \in \mathbb{B}(\gamma(t_0), \rho)$ with $p \notin \gamma[t - 1/M, t + 1/M]$ then $V_\ell(p) \neq 0$ for some $j \leq i - 1$.*

Proof. Let $\delta = M \cdot \rho$ and assume the that $V_\ell(p) = 0$ for all $j \leq i - 1$. By Lemma 14 we get that there exists $t^* \in [t - 1/M, t + 1/M]$ such that

1. $\|\gamma(t_0 + t^*) - \gamma(t_0)\|_2 \leq \frac{\delta}{M}$.
2. $D^\top(\gamma(t^* + t_0)) \cdot (\gamma(t^* + t_0) - p) = 0$.

Using the fact that the matrix $\Phi\left(\gamma(t + t_0)\right)$ admits singular value greater than σ_{min} we get that,

1. $\|\gamma(t_0 + t^*) - \gamma(t_0)\|_2 \leq \frac{\delta}{M}$.
2. $p = \gamma(t_0 + t^*) + \sum_{j=1}^{i-1} \mu_j \cdot \Phi_j\left(\gamma(t_0 + t^*)\right)$.

By the fact that $\|\gamma(t_0) - p\|_2 \leq \frac{\delta}{M}$ (recall that $\delta = M \cdot \rho$) we get that,

$$\|\sum_{j=1}^{i-1} \mu_j \cdot \Phi_j\left(\gamma(t_0 + t^*)\right)\|_2 = \|\gamma(t_0 + t^*) - p\|_2 \leq \|\gamma(t_0) - \gamma(t_0 + t^*)\|_2 + \|\gamma(t_0) - p\|_2 \leq 2\frac{\delta}{M}$$

and thus

$$\|\mu\|_2 \leq \frac{2\delta}{\sigma_{\text{min}} \cdot M} \tag{13}$$

Recall that $\|\Phi_j(x) - \Phi_j(y)\|_2 \leq L \cdot \|x - y\|_2$ and thus by applying the Taylor expansion on $V_j(\cdot)$ we get that

$$\left| V_j(p) - V_j\left(\gamma(t_0 + t')\right) - \Phi_\ell^\top\left(\gamma(t + t^*)\right) \cdot \sum_{j=1}^{i-1} \mu_j \Phi_j\left(\gamma(t + t^*)\right) \right| \leq \Theta\left(L \cdot \|\gamma(t + t_0) - p\|_2^2\right)$$

Since $V_\ell(p) = V_\ell\left(\gamma(t + t^*)\right) = 0$

$$\left| \Phi_\ell^\top\left(\gamma(t + t^*)\right) \cdot \sum_{j=1}^{i-1} \mu_j \Phi_j\left(\gamma(t + t^*)\right) \right| \leq \Theta\left(L \cdot \|\sum_{j=1}^{i-1} \mu_j \cdot \Phi_j\left(\gamma(t + t^*)\right)\|^2\right) \leq \Theta\left(L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|_2^2\right)$$

meaning that $|\left[\Phi^T \cdot \Phi \cdot \mu\right]_\ell| \leq \Theta\left(L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|_2^2\right)$ and thus

$$\sigma_{\text{min}}^2 \|\mu\|_2 \leq \|V^T \cdot V\mu\|_2 \leq \Theta\left(\sqrt{n} \cdot L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|_2^2\right) \rightarrow \|\mu\|_2 \geq \Theta\left(\frac{\sigma_{\text{min}}^2}{\sqrt{n} \cdot L \cdot \sigma_{\text{max}}^2}\right)$$

selecting $\delta \geq \Theta\left(\frac{\sigma_{\text{min}}^3 \cdot M}{\sqrt{n} \cdot L \cdot \sigma_{\text{max}}^2}\right)$ leads to contradiction. □

We conclude the section with the proof of Lemma 9. Let $\text{Vol}^n(\rho)$ denote the volume of n-dimensional ball with radius ρ and let us assume that $\gamma(t) \in [0,1]^n$ for all $t \in (0, \frac{2 \cdot 2^n}{M \cdot \text{Vol}^n(\rho/2)}]$ where $\rho = \Theta\left(\frac{\sigma_{\text{min}}^3}{\sqrt{n} \cdot \sigma_{\text{max}}^2 \cdot L}\right)$.

Let $t_i := t_1 + \frac{2i}{M}$ and let the ball $\mathbb{B}_i := \mathbb{B}(\gamma(t_i), \rho/2)$ where $\rho = \Theta\left(\frac{\sigma_{\text{min}}^3}{\sqrt{n} \cdot \sigma_{\text{max}}^2 \cdot L}\right)$. Thus there are $\frac{2^n}{\text{Vol}^n(\rho/2)}$ such balls. Notice that by Lemma 15, $\mathbb{B}_i \cap \mathbb{B}_j = \varnothing$ for $i \neq j$ and thus the latter is a contradiction due to the fact that $\mathbb{B}_i \cap [0,1]^i$ are disjoint sets with volume greater than $\frac{\text{Vol}^n(\rho/2)}{2^n}$.

G.6 PROOF OF LEMMA 10

Let $Z^\ell(x) = \{\text{coordinates } j \leq \ell - 1 \text{ such that } V_j(x) = 0\}$ and $F^\ell(x) = \{\text{coordinates } j \leq \ell - 1 \text{ that are satisfied at } x\}$
By the definition of pivot we known that $V_\ell(x) > 0$. In case $x_\ell \in (0,1)$ then the coordinate is not frozen and the first item of Lemma 10 follows. In case $x_\ell = 0$ and $[D_{Z^\ell(x)}(x)]_\ell \geq 0$ then again the first item follows. As a result, without loss of generality we assume that $[D_{Z^\ell(x)}(x)]_\ell < 0$ and $x_\ell = 0$.

At first notice that in case $Z^\ell(x) = \varnothing$ then by Definition 2 we get that $[D^\ell(x)]_\ell = 1$ which contradicts with the fact that coordinate ℓ is frozen. Also notice that since $x_\ell = 0$, Assumption 2 implies that $x_j \in (0,1)$ for all coordinates $j \in Z^\ell(x)$.

Let assume that $x_{\ell-1} = 0$. As discussed above, Assumption 2 implies that $\ell - 1 \notin Z^\ell(x)$ and thus $Z^{\ell-1}(x) = Z^\ell(x)$ which implies that $\text{sign}\left([D^{\ell-1}(x)]_{\ell-1}\right) = \text{sign}\left([D^\ell(x)]_\ell\right)$ and thus coordinate $\ell - 1$ is also frozen. As a result, the only candidate is the coordinate

$$i := \text{ the maximum } k \leq \ell \text{ with } x_k > 0$$

Note the existence of such a coordinate is guaranteed by the fact that $Z^\ell(x) \neq \varnothing$ and by the fact that for all $j \in Z^\ell(x)$, $x_j \in (0,1)$ (Assumption 2).

Let us consider the case where $x_i = 1$. Notice again that by Assumption 2, $i \notin Z^{i+1}(x) = Z^\ell(x)$ and thus $Z^i(x) = Z^{i+1}(x) = Z^\ell(x)$ which implies that $[D^i(x)]_i < 0$. Thus coordinate i is not frozen and at the same time $V_i(x) > 0$ since coordinate i is satisfied at x and $V_i(x) \neq 0$ (Assumption 2).

Now let us consider the case where $x_i \in (0,1)$ and coordinate i is not frozen. Due to the fact that x is a pivot and thus coordinate i is satisfied, we get that $V_i(x) = 0$ and thus $i \in Z^\ell(x)$. Let $D^\ell(x) := (d_1, \ldots, d_i, d_\ell)$ and $D^i(x) := (\hat{d}_1, \ldots, \hat{d}_i)$. Let us assume that $|Z^\ell(x)|$ is even. Then by Definition 2 we get that,

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_i(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_i(x)}{\partial x_i} & d_i \\ \frac{\partial V_1(x)}{\partial x_\ell} & \ldots & \frac{\partial V_i(x)}{\partial x_\ell} & d_\ell \end{vmatrix} > 0 \tag{14}$$

Since $d_\ell < 0$ then we get that

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_i(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_i(x)}{\partial x_i} & d_i \\ 0 & \ldots & 0 & d_1^2 + \ldots + d_i^2 + d_\ell^2 \end{vmatrix} < 0 \tag{15}$$

implying that

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_i(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_i(x)}{\partial x_i} \end{vmatrix} < 0 \tag{16}$$

Since $|Z^\ell(x)|$ is even then $|Z^i(x)|$ is odd ($Z^\ell(x) = Z^i(x) \cup \{i\}$) and thus by Definition 2

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \hat{d}_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \hat{d}_i \end{vmatrix} < 0 \tag{17}$$

Multiplying Equation 17 with Equation 15 we get that,

$$(\hat{d}_1, \ldots, \hat{d}_i)^\top \cdot \left(\frac{\partial V_i(x)}{\partial x_1}, \ldots, \frac{\partial V_i(x)}{\partial x_i} \right) > 0$$

H PROOF OF LEMMA 5

Let the pivots x_1 and x_2 with admissible pairs (i, S) and (i', S') respectively. Consider the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x_1$ and the trajectory $\dot{y}(t) = D^{i'}_{S'}(y(t))$ for $y(0) = x_2$ where $x_1 \neq x_2$.

We first assume that $x^* = z(t_1)$ for some t_1 and $x^* = y(t_2)$ for some t_2 where $\dot{y}(t) = D^{i'}_{S'}(y(t))$ for $y(0) = x_2$ and we will reach a contradiction. Let $M := (S'/S) \cup (S/S')$.

- $|M| \geq 2$:
  - $i = i'$: In this case $i \notin S'$ and $i' \notin S$. Let $\ell_1, \ell_2 \in M$. We will show that ℓ_1 (resp. ℓ_2) lies on the boundary ($x_{\ell_1} = 0$ or $x_{\ell_1} = 0$) and $V_{\ell_1}(x^*) = 0$. Once the latter is established, consider the set of coordinates $A := S \cup S' \cup \{i\} \cup \{i'\}$. Notice that $x^*_j = 0$ or $x^*_j = 1$ for any coordinates $j \notin A$. At the same time there exist two coordinates $\ell_1, \ell_2 \in A$ that both lie on the boundary and admit $V_{\ell_1}(x^*) = V_{\ell_2}(x^*) = 0$. The latter violates Assumption 2.

    Up next we establish that $x^*_{\ell_1} = 0$ and $V_{\ell_1}(x^*) = 0$. Without loss of generality let $\ell_1 \in S'/S$. Since $x' = \text{Next}(x_2)$ and $\ell_1 \in S'$ then Lemma 4 implies that $V_{\ell_1}(x^*) = 0$. At the same time since $\ell_1 \notin S$ and $\ell_1 \neq i$, we get that either $x^1_{\ell_1} = 0$ or $x^1_{\ell_1} = 1$. Since $\ell_1 \notin S$ the coordinate ℓ_1 stands still in the trajectory $\dot{z}(t) := D^i_S(z(t))$ with $z(0) = x_1$ and thus $x^*_{\ell_1} = 1$ or $x^*_{\ell_1} = 0$.

  - $i' > i$: Since x_1 is a pivot at which i is the under examination coordinate. By Definition 4 we get that $x^1_{i'} = 0$. The latter implies that $x^*_{i'} = 0$ since coordinate i' stands still in the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x_1$. Consider the set $A := \{j \leq i' - 1 :\ V_j(x^*) = 0\}$. Since $x^*_j = 0$ or $x^*_j = 1$ for all $j \notin A \cup \{i'\}$ and $x^*_{i+1} = 0$, Assumption 2 implies that $x^*_j \in (0,1)$ for all coordinates $j \in A$. Then Definition 5 and Definition 7 imply that $S = \{j \leq i - 1 :\ V_j(x^*) = 0\}$.

    Since (i, S) is the admissible pair of pivot x_1, $x^1_j = 0$ for $j \geq i+1$ which implies that $x^*_j = 0$ for all $j \geq i+1$. As a result, $V_j(x^*) \neq 0$ for all $j \geq i+1$. Then Definition 5 and Definition 7 imply that $S' \subseteq \{j \leq i-1 :\ V_j(x^*)\} \cup \{i\} = S \cup \{i\}$. However the latter contradicts with the fact that $|M| = 2$.

- $|M| = 1$ and $i' > i$: Without loss of generality we consider $\ell \in S'/S$. Since $\ell \in S'$, by Lemma 4 we get that $V_\ell(x^*) = 0$. At the same time since i is the under examination coordinate at x_1 and $i' > i$ by Lemma 4 we get that $x^*_{i'} = 0$.
  - $\ell \neq i$: Since $\ell \neq i$ and $\ell \notin S$, we get that either $x^1_\ell = 0$ or $x^1_\ell = 1$. Since coordinate ℓ stands still in the trajectory $\dot{z}(t) := D^i_S(z(t))$ with $z(0) = x_1$ we get that $x^*_\ell = 0$ or $x^*_\ell = 1$.

    Since $x^* = \text{Next}(x_1)$ and i is the under examination variable at x_1, by Lemma 4 we know that $x^*_j = 0$ for all $j \geq i+1$. As a result, $x^*_{i'} = 0$. Now consider the set of coordinates $A = \{j \leq i' - 1 :\ V_j(x^*) = 0\} \cup \{i'\}$. Notice that $x^*_j = 0$ or $x^*_j = 1$ for all coordinates $j \notin A$. At the same time both coordinates $i', \ell \in A$ lie on the boundary at x^*. The latter contradicts with Assumption 2.

  - $\ell = i$: In this case $S' = S \cup \{i\}$. Since $i' > i$ and i is the under examination coordinate at point x_1 we get that $x^*_{i'} = 0$. Due to the fact that i' is the under examination coordinate at x_2 we also get that $[D^{i'}_{S'}(x^*)]_{i'} = [D^{i'}_{S \cup \{i\}}(x^*)]_{i'} \leq 0$ while Assumption 3 implies $[D^{i'}_{S'}(x^*)]_{i'} = [D^{i'}_{S \cup \{i\}}(x^*)]_{i'} < 0$. The latter implies that $D^i_S(x^*)^\top \cdot \nabla V_i(x^*) > 0$[3].

    Since $i \in S'$, Lemma 4 implies that $V_i(x^*) = 0$. Since i is the under examination coordinate at x_1, by Lemma 10 we that $V_i(z(t)) > 0$ for all $t \in (0, \delta)$ where δ is sufficiently small. Since $V_i(x^*) = 0$ the latter implies that $D^i_S(x^*)^\top \nabla V_i(x^*) \leq 0$ which is a contradiction.

- $|M| = 1$ and $i' = i$: Consider $\ell \in S'/S$. Since $x^* = \text{Next}(x_2)$ by Lemma 4 we get that $V_\ell(x^*) = 0$ since $\ell \in S'$. By the fact that $\ell \notin S$, $i \neq \ell$ and x_1 is a pivot, we know by Definition 4 that either $x^1_\ell = 0$ or $x^1_\ell = 1$. Since $[D^i_S(\cdot)]_\ell = 0$, we get that $x^*_\ell = 0$ or $x^*_\ell = 1$.

  Let us consider the following mutually exclusive cases,
  - $x_i = 0$ or $x_i = 1$: Consider the set of coordinates $A = \{j \leq i - 1 :\ V_j(x^*) = 0\} \cup \{i\}$. For all coordinates $j \in S$ it holds $V_j(x^*) = 0$ while for $j \notin S$, $x^*_j = 0$ or $x^*_j = 1$. Since $S \subseteq A$, all coordinates $j \notin A$ admit $x^*_j = 0$ or $x^*_j = 1$. Notice that both coordinates $i, \ell \in A$ lie on the boundary at x^* which contradicts Assumption 2.
  - $x_i \in (0,1)$ and $V_i(x^*) = 0$: Consider the set $A := \{j \leq i - 1 :\ V_j(x^*) = 0\} \cup \{i\} \cup \{i+1\}$. For all coordinates $j \in S$ it holds $V_j(x^*) = 0$ while for $j \notin S$, $x^*_j = 0$ or $x^*_j = 1$. Since $S \subseteq A$, all coordinates $j \notin A$ admit $x^*_j = 0$ or $x^*_j = 1$. At the same time coordinates $\ell, i+1 \in A'$ lie on the boundary at x^*. The latter violates Assumption 2.
  - $x_i \in (0,1)$ and $V_i(x^*) > 0$: Without loss of generality we assume that $x^*_\ell = 0$. By Lemma 4 we know that ℓ remains satisfied during the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x_1$. Moreover $z_\ell(t) = 0$ since $[D^\ell_S(z(t))]_\ell = 0$ and $x^1_\ell = 0$. The latter implies that $D^i_S(x^*)^\top \cdot \nabla V_\ell(x^*) \geq 0$.

    Since $\ell \in S'$, we get that $y_\ell(t) \in (0,1)$ for $t \in (0, t_2)$. The latter implies that $[D^i_{S'}(x^*)]_\ell = [D^i_{S \cup \{\ell\}}(x^*)]_\ell \leq 0$. By Assumption 3 we additionally get that $[D^i_{S'}(x^*)]_\ell = [D^i_{S \cup \{\ell\}}(x^*)]_\ell < 0$. Then Lemma 12 implies that $D^i_S(x^*)^\top \cdot \nabla V_\ell(x^*) < 0$ which is a contradiction.

---

[3] See Equations (14)-(17) in the proof of Lemma 10.

- $|M| = 0$ and $i' > i$: Without loss of generality let us assume that $i' > i$. Since i is the under examination variable at x_1, Lemma 4 implies that $x^*_{i'} = 0$. Since $y_{i'}(t) \in [0,1]$ for all $t \in [0, t_2]$ we additionally get that $[D^{i'}_S(x^*)]_{i'} < 0$. Since $i \notin S$, $\text{sign}([D^i_S(x^*)]_i) = \text{sign}([D^{i'}_S(x^*)]_{i'}) < 0$ (see the proof of Lemma 10).

  Since $i < i'$ we know that coordinate i is satisfied at x^* and thus one of the following holds,

  - $x^*_i \in (0,1)$ and $V_i(x^*) = 0$.
  - $x^*_i = 1$ and $V_i(x^*) \geq 0$.
  - $x^*_i = 0$ and $V_i(x^*) \leq 0$.

  Since $i \notin S$ coordinate i lies on the boundary at point x_2 while it stands still in the trajectory $\dot{y}(t) = D^{i'}_S(y(t))$ with $y(0) = x_2$. Thus $x^*_i = 0$ or $x^*_i = 1$. The latter excludes the first case. Since coordinate i is under examination at x_1, in case $x^*_i = 1$ then $[D^i_S(x^*)]_i \geq 0$ which contradicts with the fact that $[D^i_S(x^*)]_i < 0$. Up next we exclude the third case where $x^*_i = 0$ and $V_i(x^*) \leq 0$. By Lemma 10 we know that $V_i(z(t)) > 0$ for all $t \in (0, \delta)$ once δ is selected sufficiently small. The latter together with the fact that $V_i(x^*) \leq 0$ implies that $V_i(x^*) = 0$. Now consider the set $A := S \cup \{i\} \cup \{i'\}$ and notice that $x^*_j = 0$ for all $j \notin A$. The fact that $x^*_i = x^*_{i'} = 0$ and $V_j(x^*) = 0$ for all $j \in S \cup \{i\}$ contradicts with Assumption 2.

- $|M| = 0$ and $i' = i$: In case $i = i'$ then $\dot{z}(t) = D^i_S(z(t))$ and $\dot{y}(t) = D^i_S(y(t))$. The Lipschitz-continuity of $D^i_S(\cdot)$ implies that $z(t_1) = y(t_2)$ can only occur in case $x_1 = x_2$.

I PROOF OF LEMMA 6

Let the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x$ where x is a pivot and (i, S) is an admissible pair for pivot x. Let us assume that there exists t^* such that $z(t^*) = (0, \ldots, 0)$ and $z(t)$ is not a pivot for all $t \in (0, t^*)$. Notice that in case $|S| = 0$ then $[D^i_S(z(t))]_i = 1$ which leads to a contradiction. As a result, $|S| \geq 1$. Notice that for all $j \in S$, $V_j(z(t)) = 0$ for all $t \in [0, t^*]$ and thus $V_j(0, \ldots, 0) = 0$ for all coordinates $j \in S$. Now consider the set $A = \{j \leq i - 1 : \ V_j(0, \ldots, 0) = 0\} \cup \{i\}$. Notice that all coordinates $j \in A$ admit $V_j(0, \ldots, 0) = 0$ and Assumption 3 is violated.

J DISCRETE-TIME DYNAMICS

We begin with the adaptation of the Dynamics 2 to discrete-time algorithms. The main change we need to make is to change the step 5 of Dynamics 2 to the following $z^{(k+1)} \leftarrow z^{(k)} + D^i_S(z^{(k)})$. But then we need also to adapt the notion of exit points as follows.

Definition 10 ((ϵ, γ)-Exit Points)**.** *Suppose $i \in [n]$, $S \subseteq [i-1]$ and x' is a point where coordinates in S are zero-satisfied and coordinates in $[i-1] \setminus S$ are boundary-satisfied. Then x' is an* exit point *for epoch (i, S) iff it satisfies one of the following:*

- ***(Good Exit Point)**: Coordinate i is almost satisfied at x', i.e., $\|V_i(x')\| \leq \epsilon$, or $x'_i = 0$ and $V_i(x') < \epsilon$, or $x'_i = 1$ and $V_i(x') > -\epsilon$.*

- ***(Bad Exit Point)**: For some $j \in S \cup \{i\}$, it holds that $(D^i_S(x'))_j > 0$ and $x'_j = 1$, or $(D^i_S(x'))_j < 0$ and $x'_j = 0$; in other words, if the dynamics for epoch (i, S) were to continue from x' onward, they would violate the constraints.*

- ***(Middling Exit Point)**: Let $x'' = x' + \gamma D^i_S(x')$ and for some $j \in [i-1] \setminus S$, one of the following holds: $V(x''_j) > 0$ and $x'_j = 0$, or $V(x''_j) < 0$ and $x'_j = 1$; in other words, if the dynamics for epoch (i, S) were to continue from x' onward, some boundary-satisfied coordinate would become unsatisfied.*

We next present our solution concept for the discrete-time dynamics.

Definition 11. *We say that a point x is an α-approximate solution of* $\text{VI}(V, [0,1]^n)$ *if and only if $V(x)^\top (x - y) \leq \alpha$.*

Dynamics 4 Discrete STay-ON-the-Ridge with step size γ and errors α, ϵ

```
 1: Initially $x^{(0)} \leftarrow (0, \dots, 0)$, $i \leftarrow 1$, $S \leftarrow \emptyset$, $m \leftarrow 0$.
 2: while $x^{(m)}$ is not an $\alpha$-approximate VI solution do
 3:     $z^{(0)} \leftarrow x^{(m)}$
 4:     while $\Pi(z^{(k)})$ is not an $(\epsilon, \gamma)$-exit point as per Definition 10 do
 5:         $z^{(k+1)} \leftarrow z^{(k)} + \gamma \cdot D_S^i(z^{(k)})$
 6:         $k \leftarrow k + 1$
 7:     end while
 8:     $x^{(m+1)} \leftarrow \Pi(z^{(k)})$
 9:     if $x^{(m+1)}$ is a (Good Exit Point) as in Definition 10 then
10:         if $i$ is zero-satisfied at $x(t + 1$ then
11:             Update $S \leftarrow S \cup \{i\}$.
12:         end if
13:         Update $i \leftarrow i + 1$.
14:     else if $x^{(m+1)}$ is a (Bad Exit Point) as in Definition 10 for $j = i$ then
15:         Update $i \leftarrow i - 1$ and $S \leftarrow S \setminus \{i - 1\}$.
16:     else if $x^{(m+1)}$ is a (Bad Exit Point) as in Definition 10 for $j \neq i$ then
17:         Update $S \leftarrow S \setminus \{j\}$.
18:     else if $x^{(m+1)}$ is a (Middling Exit Point) as in Definition 10 for $j < i$ then
19:         Update $S \leftarrow S \cup \{j\}$.
20:     end if
21:     Set $m \leftarrow m + 1$.
22: end while
23: return$x^{(m)}$
```

We also define $\Pi : \mathbb{R}^n \to [0,1]^n$ to be the Euclidean projection of a vector in $\mathbb{R}^n$ to the hypercube $[0,1]^n$. In Dynamics 4 we define our discrete-time dynamics for which we show Theorem 2.

Theorem 2. *We assume Assumptions 1, 2, and 3. For every $\alpha > 0$, there exist constants ϵ, γ, $\bar{M}$, K such that Dynamics 4 with step size γ and error ϵ finish after $M \le \bar{M}$ iterations of the while-loop at line 2 and it holds that $x^{(M)}$ is an α-approximate solution of $\mathrm{VI}(V, [0,1]^n)$. Additionally, for every iteration $m \le M$ of the while-loop in line 2, the while-loop in line 4 does at most K iterations.*

Proof. The main idea of the proof is to show that, for sufficiently small step size γ, the Dynamics 4 will always stay in Euclidean distance at most $\delta := \alpha/\Lambda$ from the continuous-time Dynamics 2. Then, since Dynamics 2 converge to a solution of $\mathrm{VI}(V, [0,1]^n)$ (see Theorem 1) and since V is Λ-Lipschitz we conclude that the discrete Dynamics 4 will converge to a point that is an α-approximate solution of $\mathrm{VI}(V, [0,1]^n)$.

The proof of Theorem 2 boils down to showing that there exists a step size γ and an error ϵ such that Dynamics 4 are always α/Λ close to Dynamics 2. To show this we use standard tools for the error of Euler discretized differential equations. In particular we use the following theorem.

Theorem 3 (Section 1.2 of Iserles (2009))**.** *Let $y(t) \in \mathbb{R}^n$ be the solution to the differential equation $\dot{y} = G(y)$ with initial condition $y(0) = w$, where G is a Lipschitz map $\mathbb{R}^n \to \mathbb{R}^n$. Let also $y^{(k+1)} = y^{(k)} + \gamma \cdot G(y^{(k)})$, with initial condition $y^{(0)} = w'$, with $\|w - w'\|_2 \le \zeta$. Then, for every $\eta > \zeta$ and every $T > 0$, there exists a step size $\gamma > 0$ such that*

$$\|y(k \cdot \gamma) - y^{(k)}\|_2 \le \eta \quad \text{for all} \quad 0 \le k \le T/\gamma.$$

Additionally, if the above holds for some $\gamma = \bar{\gamma}$ then it also holds for all $\gamma \le \bar{\gamma}$.

Given that $D_S^i(x)$ is Lipschitz (see Lemma 11) we can apply Theorem 3 to the while-loop of line 4 in Dynamics 4 and inductively show that $x^{(m)}$ of Dynamics 4 is close to the corresponding point of Dynamics 2.

Let τ_j be the value of the τ_{exit} variable after the j-th time that the while-loop of line 4 in Dynamics 2 has ended. For every $i \in \mathbb{N}$ we define $t_i = \sum_{j=1}^{i} \tau_i$. Our goal is to show that the $\|x^{(m)} - x(t_m)\|_2$

is small. We do this inductively. For the base of our induction observe that $x^{(0)} = x(0)$. Now assume that we have chosen a step size γ_m and that we have achieved $\|x^{(m)} - x(t_m)\|_2 \leq \zeta_m$ Also we assume as an inductive hypothesis that before the beginning of mth while-loop of line 4 we have same epoch (i, S) in both the execution of Dynamics 2 and the execution of Dynamics 4. Then, in the next execution of the while-loop of line 4 we have that $\|z^{(0)} - z(0)\|_2 \leq \zeta_m$. Also, from the proof of Theorem 1 we know that there exists a finite τ_{m+1} such that $z(\tau_{m+1})$ is an exit point. Hence, we can apply Theorem 3 and we get that for every $\eta > \zeta_m$, there exists a step size Γ_{m+1} such that

$$\|z^{(k)} - z(k \cdot \Gamma_{m+1})\|_2 \leq \zeta_m + \frac{\delta}{2^{m+1}} := \zeta_{m+1} \quad \text{for all} \quad 0 \leq k \leq \tau_{m+1}/\Gamma_{m+1}.$$

Since $x(t_m + \tau_{m+1}) = z(\tau_{m+1})$ we get that $\|x^{(m+1)} - x(t_{m+1})\|_2 \leq \zeta_{m+1}$. The only thing that is missing is to show that the update on (i, S) will be the same in the continuous and the discrete dynamics. Observe that if an exit point happens in the continuous dynamics then due to the Lipschitzness of V the same exit point has to occur in as an $(\zeta_{m+1}, \Gamma_{m+1})$-exit point in the discrete dynamics. Now repeating the argument from the proof of Theorem 1 we can easily show that it is impossible for more than one exit events to happen even in the discrete case. In particular, this follows easily from Assumption 2 and Assumption 1. Hence, the update on (i, S) will be the same. Then, we set $\gamma_{m+1} = \min\{\gamma_m, \Gamma_{m+1}\}$ and due to the last sentence of Theorem 3 we know that using the step size γ_{m+1} in all the steps before $m + 1$ it will result only to better guarantees for the distance between $x^{(\ell)}$ and $x(t_\ell)$ and therefore our induction follows. At the last iteration M we will have

$$\|x^{(M)} - x(t_M)\|_2 \leq \zeta_M \leq \delta \left(\sum_{j=1}^{m} \frac{1}{2^j} \right) \leq \delta.$$

Since $x(t_M)$ is a solution to $\mathrm{VI}(V, [0,1]^n)$ we have that $x^{(M)}$ is an α-approximate solution to $\mathrm{VI}(V, [0,1]^n)$ and the step size that we used is $\gamma = \gamma_M$.

Finally, the quantities $\bar{M}$ and K are bounded by the constant $\bar{T}$ of Theorem 1 divided by $\gamma = \gamma_m$. $\square$