STay-ON-the-Ridge: Guaranteed Convergence to Min-Max Critical Points in Nonconvex-Nonconcave Games

Costantinos Daskalakis COSTSIS@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Noah Golowich NZG@MIT.EDU
Massachusetts Institute of Technology

Stratis Skoulakis EFSTRATIOS.SKOULAKIS@EPFL.CH
École Polytechnique Fédérale de Lausanne

Manolis Zampetakis MZAMPET@BERKELEY.COM
University of California Berkeley

Editors: Gergely Neu and Lorenzo Rosasco

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 *critical or stationary points* 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 a method that is guaranteed to converge to *min-max critical points* (equiv. stationary points of gradient descent/ascent) for smooth nonconvex-nonconcave objectives in case the constraint set is the hypercube, $K = [0, 1]^n$. 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 min-max critical points, 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, 1951; Adler, 2013) and Online Learning (Blackwell, 1956; Hannan, 1957; Cesa-Bianchi and 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

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

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 (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 and Panageas (2018); Mazumdar and Ratliff (2018); Rafique et al. (2018); Hamedani and Aybat (2018); Adolphs et al. (2019); Daskalakis and Panageas (2019); Liang and Stokes (2019); Gidel et al. (2019); Mokhtari et al. (2019); Abernethy et al. (2021); Golowich et al. (2020b,a); Gorbunov et al. (2022); Cai et al. (2022). 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 and Lugosi, 2006; Shalev-Shwartz, 2012; Bubeck and Cesa-Bianchi, 2012; Shalev-Shwartz and Ben-David, 2014; Hazan, 2016). Beyond the convex/concave setting, Lin et al. (2020); Kong and 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, 2023a,b) 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 (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 and Panageas, 2018; Mazumdar and Ratliff, 2018; Jin et al., 2019; Wang et al., 2019; Daskalakis et al., 2021; Mangoubi and Vishnoi, 2021). We follow the proposal of Daskalakis et al. (2021) and target the computation of (approximate) fixed points of the gradient descent/ascent map, which correspond to *min-max critical points* (see Definition 1). These points satisfy the following for $\varepsilon > 0$ and $\delta \leq \sqrt{2\varepsilon/\Lambda}$, where Λ is the smoothness of f:

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

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

Daskalakis et al. (2021) call these points *local min-max equilibria* but we use the terminology "min-max critical points" that better captures the nature of these points (see Remark 2). An important property of min-max critical points is that they are guaranteed to exist when f is smooth.

Daskalakis et al. (2020) establish that min-max critical points correspond to approximate fixed/stationary points of the *Projected Gradient Descent/Ascent* dynamics and hence their existence follows from Brouwer's fixed point theorem.

There are a number of existing approaches which would be natural to use to find a min-max critical point, 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 min-max critical point, or converge to fixed points that are not min-max critical point (Daskalakis and 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) a second order min-max critical point. These algorithms do not enjoy global convergence guarantees, and no algorithm is known with guaranteed non-local convergence to min-max critical points.

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 and Panageas (2018); Mazumdar and 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 min-max critical points, we ask the following question: *Is there any local-search algorithm which is guaranteed to converge to a min-max critical points* in the nonconvex-nonconcave setting when $\Theta = [0,1]^k$ and $\Omega = [0,1]^\ell$? (see Table 1).

1.1. Our Contribution

In this work we propose a second-order method that is guaranteed to converge to a min-max critical point in case the corresponding constraint sets are hypercubes (Theorem 9). 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 local search based method guaranteed to min-max critical point. The only other known method is brute-force grid-search (see Section 6 for comparison). 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 computing a min-max critical point 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*, 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. We present a high-level description of our algorithm in Section 3. We explain the technical challenges in establishing its convergence and introduce the main steps of our convergence proof using a topological argument in Section 4. Then in Section 5 we provide the detailed description of our algorithm and a sketch of our proof In Table 1 we illustrate our contributions in the context of the broader literature on min-max optimization.

		convex-concave	nonconvex-concave	**nonconvex-nonconcave**
(Global) Nash Eq.	guaranteed existence	**yes**$^{\square}$	**no**†	**no**†
	computational complexity	**poly-time**‡	**NP-hard***	**NP-hard***
	convergent dynamics	**many**‡	not applicable	not applicable
Min-max Critical Pts	guaranteed existence	*same as above*	**yes**$^{+}$	**yes***
	computational 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 hypercube constraints for differing complexity in their objective function. ($\square$) v. Neumann (1928) ($\dagger$) e.g., the zero-sum 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) ($\ddagger$) see e.g. Dantzig (1951); Freund and Schapire (1997); Shalev-Shwartz (2012); Cesa-Bianchi and Lugosi (2006) ($+$) see e.g. Lin et al. (2020); Kong and Monteiro (2021); Ostrovskii et al. (2021)

2. Min-Max Critical Points

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 consider the continuous mapping $V : K \to \mathbb{R}^n$ as follows:

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

Definition 1 (Min-Max Critical Points) *A point $x = (\theta, \omega) \in K$ is called a min-max critical point if and only if $x \in K$ is a solution of the variational inequality* $\mathrm{VI}(V, K)$*, i.e. satisfies*

$$V(x)^\top \cdot (x - y) \geq 0 \quad \textit{for all } y \in K.$$

Notice that a min-max critical point $x \in K$ corresponds to a *fixed point of Projected Gradient Descent/Ascent*, $x = \Pi_K[x + V(x)]$. Theorem 5.1 of Daskalakis et al. (2020) establishes that for Λ-smooth objectives, a min-max critical point satisfies (3) and (4) for any $\varepsilon > 0$ and $\delta = \sqrt{2\varepsilon/\Lambda}$.

Remark 2 *Min-Max critical points capture the first-order optimality conditions of min-max optimization in the same way that KKT points capture the first-order optimality conditions of single-objective optimization. Thus, min-max critical points lying in the interior of K correspond to points with zero gradients similarly to KKT points lying in the interior of K. However, the two notions are crucially different for points on the boundary of K. Specifically, a min-max critical point lying on ∂K corresponds to a fixed point of gradient descent/ascent while a KKT point corresponds to a fixed point of gradient descent. We remark that this difference leads to a big difference in the computational complexity of the two notions. Specifically, computing an approximate KKT point requires polynomial time while computing an approximate min-max critical point requires exponentially many gradient-calls (see Daskalakis et al. (2021)). In Appendix A, we present further comparisons between min-max critical points and other notions of local min-max optimality.*

Remark 3 *In this work we focus on the case $K = [0, 1]^n$. In this case Definition 1 admits the following simple characterization provided in Definition 4. In Appendix B, we discuss a general technique and its challenges for extending our method to general convex constraint sets.*

Definition 4 *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 5 (Proof in Appendix F) *x is a min-max critical point iff j is satisfied at x, $\forall j \in [n]$.*

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

(L-Lipschitz) $\|V(x) - V(y)\|_2 \leq L \cdot \|x - y\|_2$, for all $x, y \in [0, 1]^n$.

(Λ-smooth) $\|J(x) - J(y)\|_F \leq \Lambda \cdot \|x - y\|_2$, 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 Definition 4.

Our algorithm is initialized at $x(0) = (0, \dots, 0)$. 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 $j \geq i + 1$.
  2. $(\nabla V_j(z))^\top \cdot d = 0$, for all $j \leq i - 1$

  Notice that Item 1 ensures that $z_j = 0$ for all $j \geq i+1$ and Item 2 that $V_j(x) = 0$ for $j \leq i-1$. We note that Item 1 and 2 specify $n - 1$ constraints on n variables. We will place mild assumptions on V so that there are exactly two (opposite) unit directions satisfying Item 1 and 2 (see Assumption 1). Moreover, in Definition 7 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', then 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 $x(0) \leftarrow (0, \dots, 0)$, $i \leftarrow 1$, $t \leftarrow 0$.
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 illustrated in Lemma 6 that has been employed to show the convergence of other equilibrium computation algorithms such as the celebrated Lemke-Howson algorithm (Lemke and Howson, 1964).

Lemma 6 *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 exists a 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 6 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:

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.

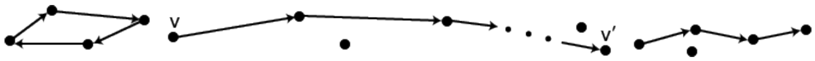


Figure 1: A directed graph whose nodes have in-degree and out-degree at most 1 is a collection of directed paths, directed cycles, and isolated nodes. Hence, if a node v has in-degree 0 and out-degree 1 then it has to be the start of a directed path that must end at a node v' after a finite number of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

3. Show that the algorithm can be initialized at some v_0 that has in-degree 0 and out-degree 1.

4. Employ Lemma 6 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 executes when it starts an epoch at such a point.

5. Detailed Description of STON'R and Main Result

We provide the formal description of our algorithm (Section 5.1), state our main convergence theorem (Section 5.2), and the main components of its proof (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 $x' \neq x$ at which 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 may hit the boundary. So, when we start a new epoch, some coordinates will be zero-satisfied and some will be boundary-satisfied. As a result, the algorithm needs to execute a continuous-time dynamics guaranteed to keep both the zero- and the boundary-satisfied satisfied. 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 (coordinates in $S \subseteq [i-1]$ 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 achieving it. In both cases, a new epoch will start. Within each epoch our algorithm executes a continuous-time dynamics that maintains all coordinates $j \in S$ zero-satisfied, all coordinates $j \in [i-1] \setminus S$ boundary-satisfied, and leaves all coordinates $j \geq i+1$ unchanged.

Definition 7 (Tangent Unit Vector of Epoch (i, S)) *Let $i \in [n]$, $S = \{s_1, \dots, s_m\} \subseteq [i-1]$, 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* $\begin{vmatrix} \frac{\partial V_{s_1}(x)}{\partial x_{s_1}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_1}} & \dots & \frac{\partial V_{s_m}(x)}{\partial x_{s_1}} & d_{s_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial V_{s_1}(x)}{\partial x_{s_m}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_m}} & \dots & \frac{\partial V_{s_m}(x)}{\partial x_{s_m}} & d_{s_m} \\ \frac{\partial V_{s_1}(x)}{\partial x_i} & \frac{\partial V_{s_2}(x)}{\partial x_i} & \dots & \frac{\partial V_{s_m}(x)}{\partial x_i} & d_i \end{vmatrix}$ *equals the sign of* $(-1)^{|S|}$.

If there is a unique unit direction satisfying the above constraints, we denote it as $D^i_S(x)$.

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 execute the dynamics $\dot{z}(\tau) = D^i_S(z(\tau))$ with $z(0) = x$, forward in time. We execute this dynamics until the earliest time τ_{exit} such that $z(\tau_{\text{exit}})$ is an *exit point* according to the next definition.

Definition 8 *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)**:* $\exists j \in S \cup \{i\}$ *such that* $\Big((D^i_S(x'))_j > 0$ *and* $x'_j = 1\Big)$, *or* $\Big((D^i_S(x'))_j < 0$ *and* $x'_j = 0\Big)$, *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 is a unique coordinate $j \le i$ triggering either a Good, a Bad or a Middling Exit Point (see Assumption 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 coordinate j triggered the bad event:
(1) if $j = i$ was responsible for the triggering, then we start epoch $(i-1, S \setminus \{i-1\})$ at x', otherwise
(2) if $j \neq i$ was responsible for the triggering, 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 D 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, $\dot{z}(\tau) = D_S^i(z(\tau))$, at $z(0) = x(t)$.
  while $z(\tau)$ is not an exit point as per Definition 8 do
    Execute $\dot{z}(\tau) = D_S^i(z(\tau))$ forward in time.
  end while
  Set $x(t + \tau) = z(\tau)$ for all $\tau \in [0, \tau_{\text{exit}}]$ (where $\tau_{\text{exit}}$ is the time $z(\tau)$ became an exit point).
  if $x(t + \tau_{\text{exit}})$ is (Good Exit Point) as in Definition 8 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 8 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 8 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 8 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 C where we present some high level reasons why they are mild.

Assumption 1 *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 $x_\ell \in \{0, 1\}\ \forall \ell \notin S \cup \{i\}$ it holds that the matrix*

$$J_S^i(x) := \begin{pmatrix} \frac{\partial V_{s_1}(x)}{\partial x_{s_1}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_1}} & \dots & \frac{\partial V_{s_m}(x)}{\partial x_{s_1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial V_{s_1}(x)}{\partial x_{s_m}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_m}} & \dots & \frac{\partial V_{s_m}(x)}{\partial x_{s_m}} \\ \frac{\partial V_{s_1}(x)}{\partial x_i} & \frac{\partial V_{s_2}(x)}{\partial x_i} & \dots & \frac{\partial V_{s_m}(x)}{\partial x_i} \end{pmatrix}$$

has singular values greater than $\sigma_{\min} > 0$ and less than $\sigma_{\max}$.

Assumption 1 ensures that for any point $x \in [0, 1]^n$ visited by Dynamics 2, the direction $D_S^i(x)$ of Definition 7 is uniquely defined (Lemma 29 in the Appendix G).

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 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 29 and Lemma 31 in the Appendix G.

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 $(d_{s_1}, \dots, d_{s_m}, d_i)$ satisfying the equations,*

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

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 $D^i_S(x)^\top V_j(x) = 0$ then higher-order information is needed in order to determine whether the direction $D^i_S(\cdot)$ makes the coordinate j satisfied or unsatisfied (see Lemma 19 in the Appendix G). We are now ready to state our main theorem.

Theorem 9 *Under Assumptions 1, 2, and 3, there exists some $\bar{T} = \bar{T}(\sigma_{\min}, \sigma_{\max}, n, L) > 0$ such that* STAY-ON-THE-RIDGE *(Dynamics 2) will stop, at some time $T \le \bar{T}$, at some point $x(T) \in [0,1]^n$ that is a min-max critical point for $K = [0,1]^n$.*

Remark 10 (Discrete-time Algorithm) *It is possible to combine the proof of Theorem 9 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 M.*

5.3. Simulated 2-Dimensional Example

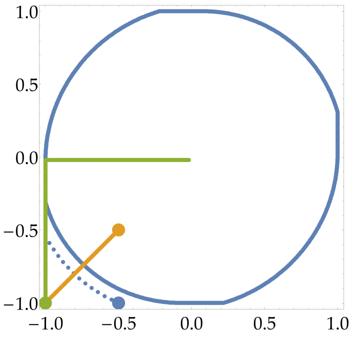


Figure 2

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 $z \le 0$, 1 for $z \ge 1$ 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 E we provide a more detailed explanation of this example and we observe similar behavior for different initializations of GDA, EG, OGDA, and FtR.

5.4. Sketch of Proof of Theorem 9

For a sketch of our proof of Theorem 9 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 G.)

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 8[1].

   Our first challenge is to show that the size of N is finite (see Lemma 17 in the Appendix G).

   Next we describe a mapping $F : N \to N$. Let $(i, S, x) \in N$, we use the dynamics $\dot{z} = D_S^i(z)$ 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 $\dot{z} = D_S^i(z)$ have a unique solution under our assumptions and hence F is well defined (see Lemma 19 in the Appendix G).

   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 Corollary 27 in the Appendix G).

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 21 in Appendix G).

3. We show that $v_0 = (1, \emptyset, (0, \ldots, 0)) \in N$ and that if we run the dynamics $\dot{z} = D_\emptyset^1(z)$ 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 $\dot{z} = D_\emptyset^1(z)$ can move forward in time and stay inside $[0,1]^n$ so v_0 has out-degree 1 (see Lemma 22 in the Appendix G).

4. The above show that our algorithm converges according to Section 4.

6. Conclusion

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 nonconvex-nonconcave 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

[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 G for the exact proof.

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 and 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.

Acknowledgements

CD was supported by NSF Awards CCF-1901292, DMS-2022448 and DMS2134108, a Simons Investigator Award, the Simons Collaboration on the Theory of Algorithmic Fairness and a DSTA grant. NG was supported by above as well as a NSF and a Hertz Foundation fellowship. MZ was supported by the Army Research Office (ARO) under contract W911NF-17-1-0304 as part of the collaboration between US DOD, UK MOD and UK Engineering and Physical Research Council (EPSRC) under the Multidisciplinary University Research Initiative (MURI).

References

Jacob D. Abernethy, Kevin A. Lai, and Andre Wibisono. Last-Iterate Convergence Rates for Min-Max Optimization: Convergence of Hamiltonian Gradient Descent and Consensus Optimization. In *the 32nd International Conference on Algorithmic Learning Theory (ALT)*, 2021.

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

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 (AISTATS)*, pages 486–495, 2019.

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

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.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *the 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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, Activity Analysis of Production and Allocation*. Wiley, New York, 1951.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *the 10th Innovations in Theoretical Computer Science (ITCS) conference*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *the 6th International Conference on Learning Representations (ICLR)*, 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 *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 *the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

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 (AISTATS)*, 2019.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *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 *the 33rd Annual Conference on Learning Theory (COLT)*, 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 *the 28th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 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 *the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 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 *the 38th International Conference on Machine Learning (ICML)*, 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 (AISTATS)*, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *the 37th International Conference on Machine Learning (ICML)*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *the 6th International Conference on Learning Representations (ICLR)*, 2018.

Oren Mangoubi and Nisheeth K Vishnoi. Greedy adversarial equilibrium: An efficient alternative to nonconvex-nonconcave min-max optimization. In *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 *the 29th 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 *the 35th International Conference on Machine Learning (ICML)*, 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 10th International Conference on Learning Representations (ICLR)*, 2022.

Thomas Pethick, Olivier Fercoq, Puya Latafat, Panagiotis Patrinos, and Volkan Cevher. Solving stochastic weak minty variational inequalities without increasing batch size. *CoRR*, abs/2302.09029, 2023a.

Thomas Pethick, Puya Latafat, Panagiotis Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. *CoRR*, abs/2302.09831, 2023b.

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.

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

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. In *the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

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

Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *the 7th International Conference on Learning Representations (ICLR)*, 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

Appendix A. Comparison with other Solution Concepts

In this section, we compare min-max critical points to other natural notions of local optimality for min-max optimization problems, furthering our discussion from Remark 2.

- ▷ **Comparison to second-order min-max critical points.** These are min-max critical points with the additional desideratum that the Hessian of the objective function f with respect to (the minimizing variables) θ has eigenvalues greater than $-\varepsilon$, i.e., $\nabla^2_\theta f(x) \succeq -\varepsilon$, and the Hessian of the objective function f with respect to (the maximizing variables) ω has eigenvalues smaller than ε, i.e., $\nabla^2_\omega f(x) \preceq \varepsilon$, where recall that $x = (\theta, \omega)$. The concept of min-max critical points that we target is strictly weaker than their second-order counterparts as they capture first-order optimality conditions and do not ask anything about the Hessian. Second-order min-max critical points are stronger in that they capture second-order optimality conditions, but their big disadvantage is that they are not guaranteed to exist even in simple cases. Furthermore, it seems likely that deciding if they exist or not is an NP-hard problem. One feature of the algorithm that we design and our proof technique is that the proof of convergence of our algorithm has as a byproduct the proof of existence of min-max critical points. Since we don't know simple conditions under which a second-order min-max critical point is guaranteed to exist we do not expect a simple adaptation of our algorithm that converges to second-order min-max critical points.

- ▷ **Comparison to local min-max equilibria.** These are points that satisfy (3) and (4) with $\varepsilon = 0$ and $\delta > 0$. In contrast, as we discussed in Section 2, min-max critical points satisfy (3) and (4) with $\epsilon > 0$ and $\delta < \sqrt{2\epsilon/L}$. Since they allow $\varepsilon > 0$, min-max critical points are strictly weaker than local min-max equilibria. The main disadvantage of local min-max equilibria is that they might not exist even in simple games; see Proposition 6 of Jin et al. (2019). Furthermore, there are no known simple conditions under which these points are guaranteed to exist and deciding if such points exist is an NP-hard problem.[2] Again, since our convergence proof has as a byproduct the proof of existence of the solutions that our algorithm targets, we do not expect a simple adaptation of our algorithm that converges to local min-max equilibria.

In sum, the reason why we do not target any of the above solution concepts is that they are not guaranteed to exist in all min-max optimization problems, while our algorithm and its associated proof of convergence using the parity argument have as a byproduct that they provide a proof of the guaranteed existence of the solution concept that they target. The same is true for algorithms using a potential function-based argument (rather than a parity argument) in their convergence proof, which are more common in optimization. The potential function-based argument proving the convergence of these algorithms also has as a byproduct a proof of the guaranteed existence of the solution concept

[2] To see this, we can start from the NP-hardness of deciding the existence of $(1/384, 1)$-local min-max equilibria shown in Theorem 10.1 of Daskalakis et al. (2020). We can then stretch the domain of these NP-hard instances and shrink the function values by considering the function $g(x) = af(bx)$, where $a = \delta$ and $b = 1/\delta$. Doing so, we reduce finding $(1/384, 1)$-local min-max equilibria of f to finding $(\delta/384, \delta)$-local min-max equilibria of g. Since finding $(0, \delta)$-local min-max equilibria is harder than finding $(\delta/384, \delta)$-local min-max equilibria we can conclude that deciding if $(0, \delta)$-local min-max equilibria exist is NP-hard for any $\delta > 0$.

that they target. In general, we cannot hope to prove guaranteed convergence of some algorithm to a particular type of solution (and only that type of solution) without at the same time providing sufficient conditions under which this type of solution is guaranteed to exist. Relatedly, note that simply *assuming* that a solution exists does not seem to help bypass this issue. The reason for this is more technical: our assumptions and the arguments in our proof, as well as potential function-based arguments, are local, i.e. they are only based on properties that hold in every small neighborhood of the domain, whereas the existence of a solution is a global property that we cannot utilize in standard potential function-based, parity, or other standard topological arguments.

Finding algorithms that are guaranteed to converge to second-order min-max critical points or local min-max equilibria is a very interesting open problem and we believe that a first step towards this is finding simple conditions under which these points exist.

Appendix B. Extension to General Convex Domains

Although our proofs only apply to the case where $K = [0, 1]^n$, there is a conceptually simple strategy to extend our method to more complex convex domains K. To implement this strategy one would need to resolve some technical difficulties that arise for different instantiations of K. Below we describe the general strategy as well as the challenges that arise in applying this strategy.

General strategy to extend STON'R to general convex domains. As discussed in Section 2, our solution concept is tantamount to a solution to a non-monotone variational inequality problem; see Definition 1. Moreover, it is a folklore fact that we can computationally efficiently reduce finding a solution to a variational inequality $\mathrm{VI}(V, K)$ over some domain of interest K to finding a solution to a variational inequality $\mathrm{VI}(V', K')$ over a different domain K', as long as *there exists an efficiently computable, Lipschitz, 1-to-1 map H from K' to K and its inverse H^{-1} is also efficiently computable and Lipschitz*. The proof of this fact goes through the equivalence of finding a solution to $\mathrm{VI}(V, K)$ and finding a fixed point of a continuous map over K. Indeed, we can define the continuous map $F : K \to K$ such that $F(x) = \Pi_K(x + V(x))$. Finding a fixed point $x \in K$ such that $F(x) = x$ is equivalent to finding a solution to $\mathrm{VI}(V, K)$; this is a well-known result that can be found, e.g. in Facchinei and Pang (2003). Next, it is easy to see that finding a fixed point of F over K can be reduced to finding a fixed point of $G(x) = H^{-1}(F(H(x)))$ over K', where $G : K' \to K'$ is a continuous map by our assumptions about H. Finally, the latter reduces to finding a solution to $\mathrm{VI}(V', K')$ where $V'(x) = G(x) - x$. Indeed, for x to be a solution to $\mathrm{VI}(V', K')$ it has to hold that $\langle V'(x), x - y\rangle \geq 0$ for all $y \in K'$. Setting $y = G(x)$ we get that $-\|x - G(x)\| \geq 0$ which only holds if $G(x) = x$. In sum, solving a VI over K reduces to solving another VI over K' assuming that H and H^{-1} exist and are Lipschitz. So, to apply STON'R to min-max optimization problems over more general convex domains K we can use $K' = [0, 1]^n$ in the above chain of reductions and ultimately run STON'R over $K' = [0, 1]^n$ and the resulting V', which will satisfy STON'R's preconditions as long as H and H^{-1} as above exist and are smooth.

Challenges. The challenge of applying the aforementioned strategy to general sets K arises from its requirements for H and H^{-1}. In particular, even for simple sets K, e.g. when K is the unit ball, the continuous 1-to-1 mappings between K and $K' = [0, 1]^n$ might not be smooth. This introduces a significant challenge because STON'R requires the smoothness of the map V to be applied. One way to bypass this issue is to define approximate mappings $\tilde{H}$ and $\tilde{H}^{-1}$ that are approximately equal to

H and H^{-1} respectively but are smooth. The design of such mappings $\tilde{H}$ and $\tilde{H}^{-1}$ is a task that may pose challenges depending on the set K. Another challenge that may arise is that the mappings H and H^{-1} might be computationally difficult to compute. We can try to bypass this issue again by designing approximate mappings that can be computed efficiently. Figuring out the details about how to design such mappings is beyond the scope of our paper and will depend on K. We believe, however, that it is important that our algorithm is not inherently restricted to hypercube constraints but there is a general strategy for applying it to more general constraint sets as long as one addresses these technical difficulties that arise.

Appendix C. Discussion Of Assumptions 1, 2 and 3

In this section we provide some heuristic arguments of why we claim that Assumptions 1, 2, and 3 are mild.

Although the arguments that we present in the next sections are heuristic, we conjecture that actually the perturbations procedures that we describe will result to a perturbed variational that will satisfy all the Assumptions 1, 2, and 3 with high probability. We leave this as an interesting open problem for follow-up work.

C.1. Assumption 1

Consider a coordinate $i \in [n]$ and the set of coordinates $S = (s_1, \ldots, s_m) \subseteq [i-1]$. Assumption 1 guarantees that for any point $x \in [0,1]^n$ such that $V_\ell(x) = 0$ for all $\ell \in S$ and $x_\ell \in \{0,1\}$ for all $\ell \notin S \cup \{i\}$, the matrix $J_S^i(x)$ admits singular values greater than $\sigma_{\min}$.

Let us try to identify all points $x \in [0,1]^n$ which violate Assumption 1, i.e, all points $x \in [0,1]^n$ for which all of the following conditions hold:

System I.

- $V_\ell(x) = 0$ for all $\ell \in S$,
- $x_\ell = 0$ for all $\ell \notin S \cup \{i\}$, and
- $\sigma_{\min}(J_S^i(x)) = 0$.

The third condition above is a strong condition since the matrix $J_S^i(x)$ has dimensions $m \times (m-1)$. As a heuristic argument to why these it is not a very strong assumption to assume that the above conditions cannot be simultaneously satisfied we argue that: (1) this system of equation is equivalent with a system with more equations than unknowns, and (2) if the map V is a map polynomials, then we can add a small random perturbation to Vso that the above system will not be satisfiable with probability 1. We start with counting the number of variables and number of equations. The above system is equivalent with finding a pair $(x, \lambda) \in [0,1]^n \times \mathbb{R}^{m-1}$ such that

System II.

- $V_\ell(x) = 0$ for all $\ell \in S$ (m equations)
- $x_\ell = 0$ for all $\ell \notin S \cup \{i\}$ ($n - m - 1$ equations)

- $\Phi_m(x) = \sum_{\ell=1}^{m-1} \lambda_\ell \cdot \Phi_\ell(x)$ where $\Phi_\ell(x) = \left(\frac{\partial V_{s_\ell}(x)}{\partial x_{s_1}}, \dots, \frac{\partial V_{s_\ell}(x)}{\partial x_{s_m}}, \frac{\partial V_{s\ell}(x)}{\partial x_i}\right)$ for $\ell \in S$ ($m+1$ equations)

As a result, such a point (x, λ) satisfies $n+m$ equations while it admits only $n+m-1$ variables.

Next, assume that V is polynomial vector field, i.e., that $V_i(x)$ is a multivariate polynomial for all $i \in [n]$. We will define $\tilde{V}(x) = V(x) + \zeta \cdot e_1$, where e_1 is first vector of the standard orthonormal basis of $\mathbb{R}^n$ and ζ is a random variable samples uniformly from the interval $[0, \varepsilon]$ where ε is a positive constant that we can choose it to be arbitrarily small. Observe that $\|\tilde{V}(x) - V(x)\|_2 \leq \varepsilon$ and hence we only lose a small approximation error when solving the variational inequality VI$(\tilde{V}, K)$ instead of VI(V, K).

So $\tilde{V}$ is still a polynomial vector field. If we interpret ζ as a constant term in System II and we homogenize the system then we get a system of $n+m$ polynomial equations with $n+m$ variables. The only way for this system to have a solution is that its resultant is 0. But the resultant of a system of polynomials is also a polynomial of the coefficients. If we fix all the other coefficient and look at the resultant as a function of the constant term ζ then the resultant is equal to some non-zero polynomial $q : \mathbb{R} \to \mathbb{R}$ on ζ which means that it has a finite number of roots. Hence if we choose ζ uniformly for an interval $[0, \varepsilon]$ we have that with probability 1 the resultant will be non-zero and hence System II cannot have any solutions.

C.2. Assumption 2

We present a heuristic argument for why Assumption 2 is mild. 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). To understand why we expect the Assumption 2 to be satisfied we define the curve $C = \{x \in [0, 1]^n$ such that $V_1(x) = 0, \dots, V_{n-1}(x) = 0\}$ and assume (because of Assumption 1) that for all $x \in C$ the matrix

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

has singular values greater than σ_{min} and smaller than σ_{max}. Assumption 2 will be violated whenever the curve C intersects an edge of the boundary of the hypercube $[\alpha_1, 1 - \beta_1] \times \cdots \times [\alpha_n, 1 - \beta_n]$. 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 hypercube $[\alpha_1, 1 - \beta_1] \times \cdots \times [\alpha_n, 1 - \beta_n]$ only in *pure facets* and not on its edges.

C.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 for $\tilde{V}$ is a $\Theta(\epsilon n)$-approximate VI solution for V.

Now Item 2 of Definition 7 defining the notion of direction $d = D^i_S(x)$ takes the following form,

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

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.

Appendix D. 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.

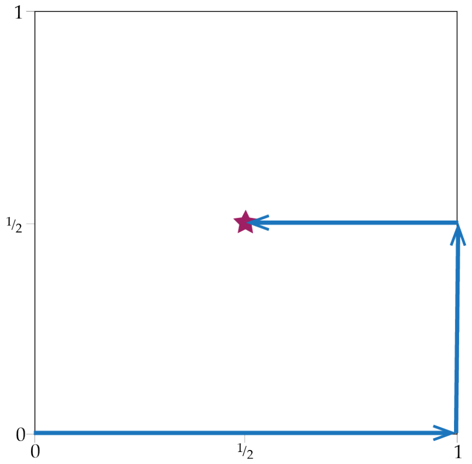


Figure 3: The path of STON'R for $f(\theta, \omega) = (\theta - 1/2) \cdot (\omega - 1/2)$.

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.

- $\triangleright$ 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 7 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 $\dot{z}(\tau) = (1, 0)$.

- $\triangleright$ 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 $x(1) = z(1) = (1, 0)$.

- $\triangleright$ 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 8, 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 $i = 2$.

- $\triangleright$ 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.

- $\triangleright$ 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 7 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 $\dot{z}(\tau) = (0, 1)$.

- $\triangleright$ 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 8 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 $x(1.5) = z(0.5) = (1, 0.5)$.

- $\triangleright$ 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 $\{1\}$.

- $\triangleright$ 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.

- $\triangleright$ 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 7 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 $\dot{z}(\tau) = (-1, 0)$.

- $\triangleright$ 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 8 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 $(\theta - 1/2) \cdot (\omega - 1/2)$.

Appendix E. 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:

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

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

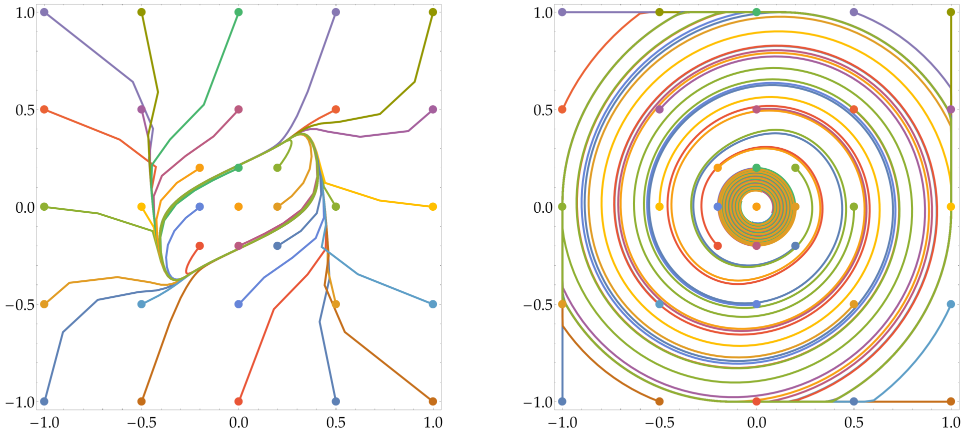
where S is the smooth-step function $S(\theta) = \begin{cases} 0, \theta \leq 0 \\ 3\theta^2 - 2\theta^3, \theta \in [0, 1] \\ 1, \theta \geq 1 \end{cases}$. Observe that in both cases it is easy to check that the only local min-max equilibrium is at $(0, 0)$.

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.

Appendix F. Proof of Lemma 5

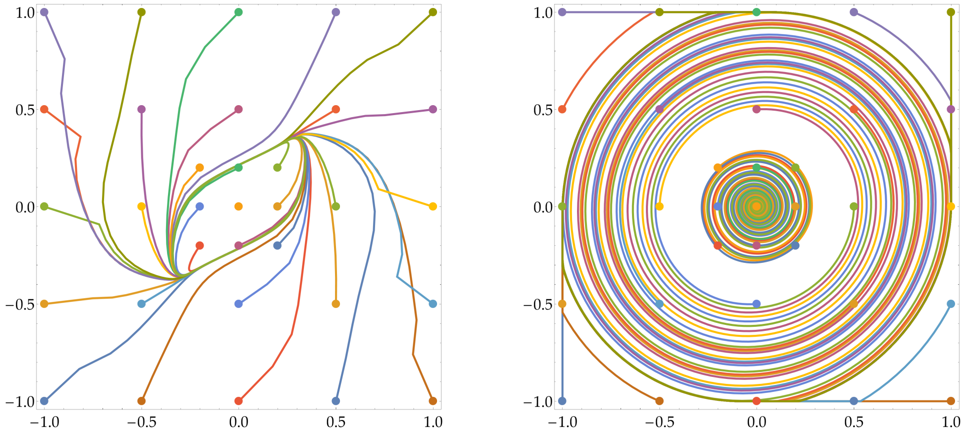
Proof ($\longleftarrow$) Let Z denote the zero-satisfied coordinates ($V_i(x) = 0$), BS^+ the boundary satisfied coordinates with $x_i = 1$ (and thus $V_i(x) > 0$) and BS^- the boundary satisfied coordinates with $x_i = 0$ (and thus $V_i(x) < 0$). For any $y \in [0, 1]^n$, we have $\sum_{i=1}^n V_i(x)(x_i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



(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.

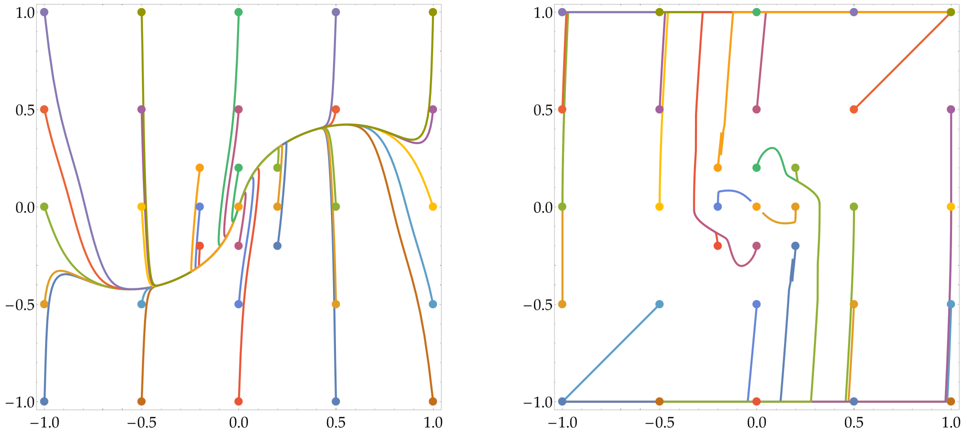


(a) $f_1(\theta, \omega)$.

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

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.

$(\longrightarrow)$ 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



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

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

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.

$V(x)^\top(x - y) \leq 0$ to get $V_i(x) \leq 0$. If $x_i \in (0, 1)$ consider first $y_i = x_i + \delta$ for some small $\delta > 0$ 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, $V_i(x) = 0$. ■

Appendix G. Proof of Theorem 9

In this section we present the proof of Theorem 9. 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.

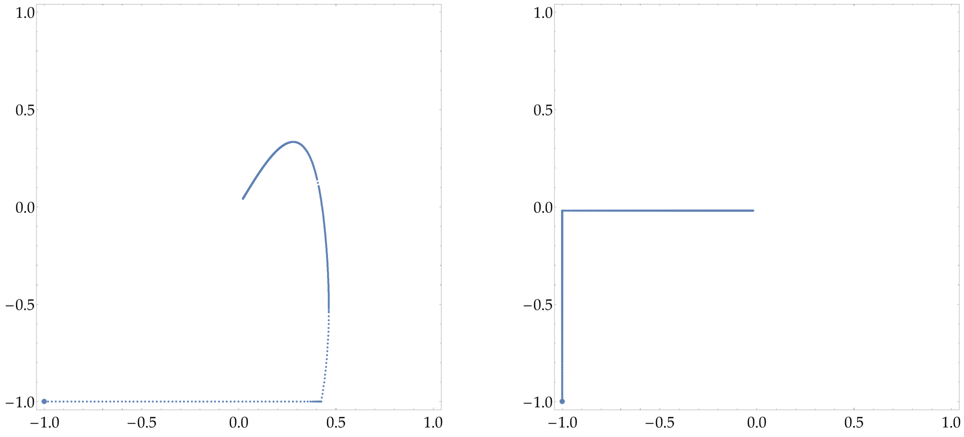
G.1. Simplifying Dynamics 2

In this section we present a more concise version of Dynamics 2. As already discussed, the proof of Theorem 9 follows the topological argument described in Section 5.4. More precisely, the "nodes" of the graph are the triples (i, S, x) where

> (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 8.

To simplify our formal arguments we introduce the notion of *pivot* $x \in [0, 1]^n$ that as we shall see in a while correspond to triples (i, S, x). From now on pivots should be thought as the nodes N.

Definition 11 *A point $x \in [0, 1]^n$ is called a pivot if and only if the following hold,*



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

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.

- *If coordinate ℓ is not satisfied then $V_\ell(x) > 0$.*
- *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 $M := \{j \leq \ell - 1 : \ V_j(x) = 0\}$.*

Corollary 12 *The point $(0, \ldots, 0) \in [0,1]^n$ is a pivot.*

Corollary 13 *Any point $x \in [0,1]^n$ at which all coordinates $\ell \in [n]$ are satisfied is a pivot.*

At step 3, Dynamics 2 follows epoch's (i, S) dynamics $\dot{z}(\tau) = D_S^i(z(\tau))$ with $z(0) = x(t)$. The point $x(t)$ was the exit point (Good, Bad or Middling) of some previous epoch (i', S') that was updated to epoch (i, S) in one of the Steps $10, 12, 14$ and 18. In Definitions 14, 15 and 16, we provide an alternative way of "locally computing" the epoch (i, S) by using $x(t)$ at Step 3 of Dynamics 2.

Definition 14 *The **ideal direction of movement** at point $x \in [0,1]^n$ with respect to coordinate $i \in [n]$, denoted as $D^i(x)$, is defined as: Let the zero-satisfied coordinates $S = \{j < i$ with $V_j(x) = 0\}$ at point $x \in [0,1]^n$ and consider the direction $D_S^i(x) := (d_1, \ldots, d_n)$ of Definition 7 (recall that $d_j = 0$ for all $j \notin S \cup \{i\}$). If for all coordinates $k \in S$, one of the following holds:*

1. $x_k \in (0,1)$

2. *if* $x_k = 0$ *then* $d_k \geq 0$
3. *if* $x_k = 1$ *then* $d_k \leq 0$

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* $D^i(x) := D^i_{S\setminus\{j\}}(x)$*. In case* $i = 0$*, we consider* $D^i(x) := D^0_{\varnothing}(x) := (0, \dots, 0)$*.*

Definition 14 captures the case where $x \in [0, 1]^n$ is Bad Exit Point for the epoch (i, S) triggered by some $j \neq i$. In case Dynamics 2 in the course of epoch (i, S), hits a point x where $\{x_j = 0$ and $d_j < 0\}$ or $\{x_j = 1$ and $d_j > 0\}$, then it updates the epoch $(i, S) \leftarrow (i, S/\{j\})$ at Step 16 and goes to Step 3 so as to start a new epoch (i, S) from $x \in [0, 1]^n$.

Definition 15 *Given a point* $x \in [0, 1]^n$ *coordinate* $i \in [n]$ *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 14).*

Definition 15 captures the case where $x \in [0, 1]^n$ is a Bad Exit Point for the epoch (i, S) triggered by $j = i$. Notice that in case Dynamics 2 in the course of epoch (i, S), hits x at which coordinate i is frozen, then it updates the epoch $(i, S) \leftarrow (i - 1, S/\{i - 1\})$ at Step 13 and starts again at Step 3 with the updated epoch (i, S) and initial point $x \in [0, 1]^n$.

In Definition 16 we present the notion of *admissible pair* (i, S) of a point $x \in [0, 1]^n$. Using this notion we can substantially simplify Dynamics 2 by extracting epoch (i, S) at Step 3 from point $x(t)$. More precisely the epoch (i, S) at Step 3 is the *admissible pair* (i, S) of $x(t)$.

Definition 16 *Given a point* $x \in [0, 1]^n$ *consider*

1. $\ell := \min_{1 \leq j \leq n}\{$*coordinate* j *is not satisfied at* $x\}$.
2. $i := \max_{j \leq \ell}\{$*coordinate* j *is not frozen at* $x\}$.
3. $S \leftarrow$ *the set of coordinates such that* $D^i(\cdot) = D^i_S(\cdot)$ *(see Definition 14).*

The coordinate i *is called the* ***under examination*** *coordinate at point* $x \in [0, 1]^n$ *and the pair* (i, S) *is called the* ***admissible pair*** *of point* $x \in [0, 1]^n$*. In case all coordinates* $j \in [n]$ *are satisfied or all coordinates* $j \leq \ell$ *are frozen then* $(i, S) := (0, \varnothing)$*.*

Notice that Steps (8)-(19) in Dynamics 2 guarantee that at epoch (i, S), all coordinates $j \leq i - 1$ are satisfied. Item 1 of Definition 16 captures the fact that coordinate i is guaranteed to be smaller than coordinate $\ell := \min_{1 \leq j \leq n}\{$coordinate j is not satisfied at $x(t)\}$. At the same time, Item 2 captures the case where $x(t)$ has been reached as a Bad Event of epoch $(S \cup \{i + 1\}, i + 1)$ triggered by $j = i + 1$ and Item 3 the case where $x(t)$ has been reached as a Bad Event of epoch $(S \cup \{j\}, i)$ triggered by $j \neq i$.

In Dynamics 3 we present a simpler description of Dynamics 2 using Definition 16.

Dynamics 3 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
    At point $x(t)$ compute the admissible pair $(i, S)$ of $x(t)$ (Definition 16).
    Follow the 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 8 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 earliest time $z(\tau)$ became an exit point).
    Set $t \leftarrow t + \tau_{\text{exit}}$.
end while
return $x(t)$
```

G.2. Proof Theorem 9 for Dynamics 3

In this section we formalize 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* of Definition 11. We first establish that the number of pivots is finite.

Lemma 17 *There exists a finite number of pivots. More precisely, there are at most* $4^n / \mathrm{Vol}^n\left(\frac{2\sigma^2_{min}}{\sqrt{n}L\sigma^2_{max}}\right)$ *where* $\mathrm{Vol}^n(\rho)$ *denotes the volume of the* n*-dimensional ball of radius* ρ*.*

As already mentioned, we encode a transition of Dynamics 3 from pivot x to pivot x' as a *directed edge* from x to x' which we denote as $x' = \mathrm{Next}(x)$. The latter is formalized in Definition 18.

Definition 18 *Let a pivot* $x \in [0,1]^n$ *with admissible pair* (i, S)*. A pivot* x' *is called next pivot of* x*,* $x' = \mathrm{Next}(x)$ *if and only if there exists a trajectory* $\dot{z}(t) = D^i_S(z(t))$ *with* $z(0) = x$ *for* $t \in [0, t^*]$ *and* $t^* > 0$ *such that*

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

In Lemma 19 we establish that any pivot $x \in [0,1]^n$ with at least one unsatisfied variable must necessarily admit out-degree 1. In other words, there exists a pivot x' with $x' = \mathrm{Next}(x)$.

Lemma 19 *For any pivot* $x \in [0,1]^n$ *with at least one unsatisfied coordinate there exists a unique pivot* x' *such that* $x' = \mathrm{Next}(x)$*. More precisely, let* (i, S) *be the admissible pair for pivot* x*, then there exists* $t^* \in [0, C]$ *(*$C > 0$ *only on* $n, \sigma_{\min}, \sigma_{\max}, L$*) such that for* $t \in [0, t^*]$ *the initial value problem*

$$\dot{z}(t) = D^i_S(z(t)) \quad \text{and} \quad z(0) = x$$

admits a unique solution. Moreover,

1. $z(t) \in [0,1]^n$ *for all* $t \in [0, t^*]$

2. *$x' := z(t')$ is a pivot for some finite $t' > 0$.*

3. *$z(t)$ is not pivot for $t \in (0, t')$.*

Additionally for all $t \in (0, t')$ the following hold,

1. *$V_j(t) = 0$ for all $j \in S$*

2. *$z_j(t) \in (0, 1)$ for all $j \in S \cup \{i\}$.*

3. *$z_j(t) \in \{0, 1\}$ for all $j \notin S \cup \{i\}$.*

4. *$z_j(t) = 0$ for all $j \geq i + 1$.*

5. *all coordinates $j \leq i - 1$ are satisfied at $z(t)$.*

Lemma 19 directly implies that any pivot $x \in [0, 1]^n$ with 0 out-degree is a solution.

Corollary 20 *Let a pivot $x \in [0, 1]^n$ such that $x' \neq \mathrm{Next}(x)$ for all pivots $x' \in [0, 1]^n$. Then all coordinates $\ell \in [n]$ are satisfied at $x \in [0, 1]^n$.*

Corollary 20 establishes that once Dynamics 3 hits a pivot with out-degree 0 then it has reached a solution. In order to guarantee that Dynamics 3 hits a pivot with 0 out-degree we need to exclude the case that Dynamics 3 cycles around pivots with at least one unsatisfied coordinate. As a first step, in Lemma 21 we establish that each pivot x admits in-degree at most 1.

Lemma 21 *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 $x_1, x_2 \in [0, 1]^n$ then $x_1 = x_2$.*

Since Dynamics 3 first visits $(0, \ldots, 0)$, Lemma 21 implies that Dynamics 3 either hits a pivot with 0 in-degree and stops or it goes over a cycle of pivots containing pivot $(0, \ldots, 0)$. The latter case implies that pivot $(0, \ldots, 0)$ admits in-degree 1. In Lemma 22 we complete the proof of Theorem 9 by establishing that the pivot $(0, \ldots, 0)$ admits 0 in-degree.

Lemma 22 *There is no pivot $x \in [0, 1]^n$ such that $\mathrm{Next}(x) = (0, \ldots, 0)$.*

Appendix H. Auxiliary Lemmas

Lemma 23 *Let a $A = (\Phi_1, \ldots, \Phi_n)$ where $\Phi_\ell \in \mathbb{R}^{n-1}$. There exists an $i \in [n]$ such that the matrix $A' = (\Phi_1, \ldots, \Phi_{i-1}, \Phi_{i+1}, \ldots, \Phi_n)$ admits singular value $\sigma_{\min}(A') \geq \sigma_{\min}(A)/\sqrt{n}$.*

Proof Notice that $A \cdot A^\top = \sum_{\ell=1}^n \Phi_\ell \cdot \Phi_\ell^\top$. Moreover there exist $\lambda_1, \ldots, \lambda_n$ such that

$$\sum_{\ell=1}^n \lambda_\ell \cdot \Phi_\ell = 0 \quad \text{where} \quad \sum_{\ell=1}^n |\lambda_\ell| \neq 0$$

Let $i := \text{argmax}_{\ell \in [n]} |\lambda_\ell|$, meaning that $\Phi_i = \sum_{\ell \neq i} \lambda_\ell \cdot \Phi_i$ where $|\lambda_\ell| \leq 1$. Notice that for any $x \in \mathbb{R}^{n-1}$ and $\|x\|_2 \leq 1$,

$$\begin{aligned}
\sigma_{\min}(A) &\leq x^\top \left(\sum_{\ell \in [n]} \Phi_\ell \cdot \Phi_\ell^\top \right) x = x^\top \left(\sum_{\ell \neq i} \Phi_\ell \cdot \Phi_\ell^\top \right) x + x^\top \left(\Phi_i \cdot \Phi_i^\top \right) x \\
&= x^\top \left(\sum_{\ell \neq i} \Phi_\ell \cdot \Phi_\ell^\top \right) x + \left(\sum_{\ell \neq i} \lambda_\ell \Phi_\ell^\top x \right)^\top \cdot \left(\sum_{\ell \neq i} \lambda_\ell \Phi_\ell^\top x \right) \\
&= x^\top \left(\sum_{\ell \neq i} \Phi_\ell \cdot \Phi_\ell^\top \right) x + 2 \sum_{\ell < \ell' \neq i} \lambda_\ell \lambda_{\ell'} \cdot x^\top \Phi_\ell \Phi_{\ell'}^\top x \\
&\leq x^\top \left(\sum_{\ell \neq i} \Phi_\ell \cdot \Phi_\ell^\top \right) x + \sum_{\ell < \ell' \neq i} \left(\frac{\lambda_\ell^2}{2} \cdot x^\top \Phi_\ell \Phi_\ell^\top x + \frac{\lambda_{\ell'}^2}{2} \cdot x^\top \Phi_{\ell'} \Phi_{\ell'}^\top x \right) \\
&\leq x^\top \left(\sum_{\ell \neq i} \Phi_\ell \cdot \Phi_\ell^\top \right) x + \sum_{\ell < \ell' \neq i} \left(\frac{1}{2} \cdot x^\top \Phi_\ell \Phi_\ell^\top x + \frac{1}{2} \cdot x^\top \Phi_{\ell'} \Phi_{\ell'}^\top x \right) \\
&\leq (n-1) \cdot x^\top \left(\sum_{\ell \neq i} \Phi_\ell \cdot \Phi_\ell^\top \right) x
\end{aligned}$$

As a result, for any $x \in \mathbb{R}^{n-1}$ with $\|x\|_2 = 1$ we get that

$$x^\top \cdot A' A'^\top \cdot x \geq \frac{\sigma_{\min(A)}}{n-1}$$

■

Lemma 24 *Given a coordinate $i \in [n]$ and a set of coordinates $S = (s_1, \ldots, s_m) \subseteq [i-1]$ consider the matrix $J_S^i(x) \in \mathbb{R}^{(m+1) \times m}$ at point $x \in \mathbb{R}^n$ is defined as follows,*

$$J_S^i(x) := \begin{pmatrix} \frac{\partial V_{s_1}(x)}{\partial x_{s_1}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_1}} & \cdots & \frac{\partial V_{s_m}(x)}{\partial x_{s_1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial V_{s_1}(x)}{\partial x_{s_m}} & \frac{\partial V_{s_2}(x)}{\partial x_{s_m}} & \cdots & \frac{\partial V_{s_m}(x)}{\partial x_{s_m}} \\ \frac{\partial V_{s_1}(x)}{\partial x_i} & \frac{\partial V_{s_2}(x)}{\partial x_i} & \cdots & \frac{\partial V_{s_m}(x)}{\partial x_i} \end{pmatrix}$$

Let a point $x \in [0,1]^n$ for which $V_j(x) = 0$ for all $j \in S$ and $x_j \in \{0,1\}$ for all $j \notin S \cup \{i\}$. Then for any $y \in \mathbb{R}^n$ such that $\|x - y\|_2 \leq \sigma_{min}/2Ln$ the singular values of $J_S^i(y)$ are greater than $\sigma_{\min}/2$,

$$\sigma_{\min}/2 \leq \sigma\left(J_S(y)\right)$$

Proof To simplify notation let $S := \{1, \ldots, i-1\}$. Since $V_j(x) = 0$ for all $j = 1, \ldots, i-1$, Assumption 1 ensures that the matrix $J_S^i(x)$ admits singular value greater than $\sigma_{\min}$. Moreover

$$\|\nabla V_j(x) - \nabla V_j(y)\|_2 \leq L \cdot \|x - y\|_2 \leq \frac{\sigma_{\min}}{2n}$$

Let $x^* := \text{argmin}_{\|x\|=1}\|J_S^i(y)\cdot x\|_2$. As a result,

$$\begin{aligned}
\sigma\left(J_S^i(y)\right) &\geq \|J_S^i(y)\cdot x^*\|_2 \\
&\geq \|J_S^i(x)\cdot x^*\|_2 - \|J_S^i(x) - J_S^i(y)\|_2 \\
&\geq \sigma_{\min} - \|J_S^i(x) - J_S^i(y)\|_F \\
&\geq \sigma_{\min} - \sqrt{n^2\frac{\sigma_{\min}^2}{4n^2}} = \frac{\sigma_{\min}}{2}
\end{aligned}$$

■

Lemma 25 *Let $x, y \in \mathbb{R}^n$, a coordinate $i \in [n]$, and a set of coordinates $S \subseteq [i-1]$ such that both matrices $J_S^i(x), J_S^i(y)$ admit singular values greater than $\sigma_{\min}$. Then,*

$$\|D_S^i(x) - D_S^i(y)\|_2 \leq M \cdot \|x - y\|_2$$

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

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

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

Since $J_S^i(x)$ admits singular values greater that $\sigma_{\min}$ then by Lemma 23 and without loss of generality

$$\sigma_{\min}(A(x)) \geq \frac{\sigma_{\min}}{\sqrt{n}}$$

Combining the latter with Lemma 24 we get that for $\|x - y\|_2 \leq \sigma_{\min}/2Ln^{3/2}$,

$$\sigma_{\min}(A(y)) \geq \frac{\sigma_{\min}}{\sqrt{n}}$$

As a result, both $A(x), A(y)$ are invertible and admit singular value greater than $\sigma_{\min}$. Moreover due to the fact that,

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

we get that

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

To simplify notation $C_A := nL$, $C_b := L$ and $\sigma_{\min} := \sigma_{\min}/\sqrt{n}$. Notice that the direction $D_S^i(x)$ of Definition 7 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 26 [3] *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 25 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 $\|x-y\|_2 \leq \frac{\sigma_{\text{min}}}{2C_A}$ implies,

$$\|F\|_2 = \|(A(x)-A(y))\cdot A^{-1}(x)\|_2 \leq \frac{C_A}{\sigma_{\text{min}}} \cdot \|x-y\|_2 \leq \frac{1}{2} \tag{5}$$

For the first case we get,

$$\|\left(\frac{A^{-1}(x)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right)\|_2^2$$
$$= \frac{\|A^{-1}(x)\cdot b(x) - A^{-1}(y)\cdot b(x)\|_2^2}{1+\|A^{-1}(x)\cdot b(x)\|_2^2}$$
$$\leq \frac{\|A^{-1}(x)\cdot b(x) - A^{-1}(y)\cdot b(x)\|_2^2}{\|A^{-1}(x)\cdot b(x)\|_2^2}$$
$$\leq \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \cdot \frac{\|F\|_2}{1-\|F\|_2}\right)^2 \qquad \text{by Lemma 26}$$
$$\leq \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \cdot 2 \cdot \|F\|_2\right)^2 \qquad \text{by Equation 5}$$
$$\leq \frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} \cdot 4 \cdot \|(A(x)-A(y))\cdot A^{-1}(x)\|_2^2 \leq 4\frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^4} \cdot C_A^2 \cdot \|x-y\|_2^2$$

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

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

For the second case

$$\begin{aligned}
&\|\left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2}}\right) - \left(\frac{A^{-1}(y)\cdot b(x)}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}, \frac{1}{\sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}}\right)\|_2^2 \\
&= \left(\|A^{-1}(y)\cdot b(x)\|^2 + 1\right) \frac{\left(\sqrt{1+\|A^{-1}(x)\cdot b(x)\|_2^2} - \sqrt{1+\|A^{-1}(y)\cdot b(x)\|_2^2}\right)^2}{(1+\|A^{-1}(y)\cdot b(x)\|_2^2)\cdot(1+\|A^{-1}(x)\cdot b(x)\|_2^2)} \\
&\leq \left(\|A^{-1}(y)\cdot b(x)\|^2 + 1\right) \frac{\left(\|A^{-1}(x)\cdot b(x)\|_2 - \|A^{-1}(y)\cdot b(x)\|\right)^2}{(1+\|A^{-1}(y)\cdot b(x)\|_2^2)\cdot(1+\|A^{-1}(x)\cdot b(x)\|_2^2)} \qquad \text{since } \sqrt{1+b} - \sqrt{1+a} \leq \sqrt{b} - \sqrt{a} \\
&\leq \frac{\left(\|A^{-1}(x)\cdot b(x)\|_2 - \|A^{-1}(y)\cdot b(x)\|\right)^2}{\|A^{-1}(x)\cdot b(x)\|_2^2} \\
&\leq \frac{\|A^{-1}(x)\cdot b(x) - A^{-1}(y)\cdot b(x)\|_2^2}{\|A^{-1}(x)\cdot b(x)\|_2^2} \\
&\leq \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \cdot \frac{\|F\|_2}{1-\|F\|_2}\right)^2 \qquad \text{by Lemma 26} \\
&\leq \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \cdot 2 \cdot \|F\|_2\right)^2 \qquad \text{by Equation 5} \\
&\leq \frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} \cdot 4 \cdot \|(A(x) - A(y)) \cdot A^{-1}(x)\|_2^2 \leq 4\frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^4} \cdot C_A^2 \cdot \|x-y\|_2^2
\end{aligned}$$

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,

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

For the forth case,

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

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

Lemma 27 *Let $x \in [0,1]^n$ and a set of coordinates $Z \subseteq \{\ell \leq i-1 \ : \ V_\ell(x) = 0\}$. Then for any coordinate $j \in Z$ the following hold,*

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

2. *If $x_j = 1$ and $[D_Z^i(x)]_j > 0$ then $\left(D_{Z/\{j\}}^i(x)\right)^\top \cdot \nabla V_j(x) > 0$.*

Proof To simplify notation let $Z = \{1, \dots, i-1\}$, $D^i_Z(x) = (d_1, \dots, d_j, \dots, d_i)$ and $D^i_{Z/\{j\}}(x) = (\hat{d}_1, \dots, \hat{d}_{j-1}, \hat{d}_{j-1}, \dots, \hat{d}_i)$. 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

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

Since i is even we get by Definition 7,

$$\begin{vmatrix}
\frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \hat{d}_1 \\
\vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial V_1(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & \hat{d}_{j-1} \\
\frac{\partial V_1(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & \hat{d}_{j+1} \\
\vdots & & \vdots & \vdots & \vdots & & \vdots \\
\frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \hat{d}_i
\end{vmatrix} > 0 \tag{6}$$

and that

$$\begin{vmatrix}
\frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_j(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_j(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & d_{j-1} \\
\frac{\partial V_1(x)}{\partial x_j} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_j} & \frac{\partial V_j(x)}{\partial x_j} & \frac{\partial V_{j+1}(x)}{\partial x_j} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_j} & d_j \\
\frac{\partial V_1(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_j(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & d_{j+1} \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_j(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & d_i
\end{vmatrix} < 0 \tag{7}$$

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 7) with $d_j < 0$ (we have assumed that $x_j = 0$) we get by Equation 7,

$$\begin{vmatrix}
\frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_1} & \frac{\partial V_j(x)}{\partial x_1} & \frac{\partial V_{j+1}(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j-1}} & \frac{\partial V_j(x)}{\partial x_{j-1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j-1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j-1}} & d_{j-1} \\
0 & \dots & 0 & 0 & 0 & \dots & 0 & d_1^2 + \ldots + d_i^2 \\
\frac{\partial V_1(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_{j+1}} & \frac{\partial V_j(x)}{\partial x_{j+1}} & \frac{\partial V_{j+1}(x)}{\partial x_{j+1}} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_{j+1}} & d_{j+1} \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{j-1}(x)}{\partial x_i} & \frac{\partial V_j(x)}{\partial x_i} & \frac{\partial V_{j+1}(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & d_i
\end{vmatrix} > 0 \tag{8}$$

which implies that

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

Multiplying with Equation 6 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 7) 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 $\Phi_\ell = \left(\frac{\partial V_\ell(x)}{\partial x_1}, \dots, \frac{\partial V_\ell(x)}{\partial x_{j-1}}, \frac{\partial V_\ell(x)}{\partial x_{j+1}}, \dots, \frac{\partial V_\ell(x)}{\partial x_i}\right)$. ■

x

Lemma 28 *Let the functions $F_1(x), \dots, F_i(x)$ where $F_\ell : [0,1]^i \mapsto \mathbb{R}$ and the set $B := \{x \in [0,1]^i : \ F_\ell(x) = 0 \text{ for all } \ell = 1, \dots, i\}$. In case $F_1, \dots, F_\ell$ satisfy the following assumptions*

1. $\|\nabla F_\ell(x) - \nabla F_\ell(y)\|_2 \leq L \cdot \|x - y\|_2$
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, $|B| \leq 2^i/\mathrm{Vol}^i\left(\frac{2\sigma_{min}^2}{\sqrt{i}L\sigma_{max}^2}\right)$ where $\mathrm{Vol}^i(\rho)$ is the volume of the i-dimensional ball with radius ρ.

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

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

which implies that

$$\|\mu\|_2 \leq \frac{\rho}{\sigma_{\text{min}}} \tag{10}$$

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

$$\|\mu\|_2 \geq \frac{2\sigma_{\text{min}}}{\sqrt{i}L\sigma_{\text{max}}^2} \tag{11}$$

Combining Equation 10 and 11 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,

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

■

Lemma 29 *Let a pivot $x \in [0,1]^n$ and (i, S) the admissible pair for pivot x. In case the under examination variable $i \geq 1$, then there exists a $t^* > 0$ such that for all $t \in [0, t^*]$ the initial value problem*

$$\dot{z}(t) = D_S^i\left(z(t)\right) \quad \textit{with} \quad z(0) = x$$

admits a unique solution. Moreover for all $t \in (0, t^*]$ the following hold,

1. *$V_j(z(t)) = 0$ for all coordinates $j \in S$.*

2. *$z_j(t) \in (0,1)$ for all coordinates $j \in S \cup \{i\}$.*

3. *($z_j(t) = 0$ and $V_j(z(t)) < 0$) or ($z_j(t) = 1$ and $V_j(z(t)) > 0$) for all $j \in [i-1]/S$.*

4. *$z_j(t) = 0$ for all coordinates $j \geq i + 1$.*

Proof Let pivot $x \in [0,1]^n$ and consider the ball $\mathcal{B} := \mathcal{B}(x, \rho)$ with $\rho = \sigma_{\text{min}}/2Ln$. Since pivot $x \in [0,1]^n$ satisfies $V_j(x) = 0$ for all $j \in S$, by Assumption 1 we know that the matrix $J_S(x)$ admits singular values greater than σ_{min}. Moreover by Lemma 24 for all $x' \in \mathcal{B}$ the matrix $J_S(x')$ admits singular values greater than $\sigma_{\text{min}}/2$. Thus by Lemma 25, for $\hat{x}, x' \in \mathcal{B}$,

$$D_S^i(\hat{x}) - D_S^i(x') \leq M \cdot \|\hat{x} - x'\|_2$$

where $M := \Theta\left(\frac{\sigma_{\text{max}}}{\sigma^2_{\text{min}}} \cdot n \cdot L\right)$. Since $D_S^i(\cdot)$ is M-Lipschitz continuous in $\mathcal{B}$ and $\|D_S^i(\cdot)\|_2 = 1$, the Picard–Lindelöf theorem implies that there exits $t^* > 0$ such that for $t \in [0, t^*]$ the initial value problem

$$\dot{z}(t) = D_S^i(z(t)) \text{ and } z(0) = x$$

admits a unique solution.

Let $Z := \{\ell \leq i-1 \ : \ V_\ell(x) = 0\}$ and $F := \{\ell \leq i-1 \ : \ \text{coordinate } \ell \text{ is boundary-satisfied at } x\}$. Since x is a pivot by Definition 11 we get that all coordinates $\ell \leq i-1$ are satisfied and thus each coordinate $\ell \leq i-1$ either belongs to Z or to F. By the Definition 11 we know that there exists at least one coordinate $\ell \in Z \cup \{i\}$ such that $x_\ell \in \{0,1\}$. As a result, by Assumption 2 we get that for all coordinates $\ell \in F$,

1. If $x_\ell = 0$ then $V_\ell(x) < 0$
2. If $x_\ell = 1$ then $V_\ell(x) > 0$

Let us first consider the case where for each coordinate $\ell \in Z$ one of the following holds,

1. $x_\ell \in (0,1)$
2. $x_\ell = 0$ and $[D_Z^i(x)]_\ell \geq 0$
3. $x_\ell = 1$ and $[D_Z^i(x)]_\ell \leq 0$

In this case by Definition 14 and Definition 16 we get that $S = Z$. Moreover by Assumption 3 we additionally get that for each coordinate $\ell \in S$ one of the following holds,

1. $x_\ell \in (0,1)$
2. $x_\ell = 0$ and $[D_Z^i(x)]_\ell > 0$
3. $x_\ell = 1$ and $[D_Z^i(x)]_\ell < 0$

Recall that by Definition 7, $\nabla V_\ell(z(t))^\top \cdot D_S^i(z(t)) = 0$ for all $\ell \in S$ and thus for all $t \in [0, \delta]$ with $\delta > 0$ being sufficiently small,

1. $V_\ell\left(z(t)\right) = 0$ for all $\ell \in S$
2. $z_\ell(t) \in (0,1)$ for all $\ell \in S \cup \{i\}$.
3. ($z_\ell(t) = 0$ and $V_\ell(z(t)) < 0$) or ($z_\ell(t) = 1$ and $V_\ell(z(t)) > 0$) for all $\ell \in F$.

4. $z_\ell(t) = 0$ for all coordinates $\ell \geq i+1$.

Now consider the case in which there exists $j \in Z$ with ($x_j = 0$ and $[D^i_Z(x)]_j < 0$) or ($x_j = 1$ and $[D^i_Z(x)]_j > 0$). By Assumption 2 we know that such a coordinate must be unique. In this case by Definition 14 we get $D^i(x) = D^i_{Z/\{j\}}(x)$ and thus by Definition 16, $S = Z/\{j\}$. Without loss of generality let assume that

$$x_j = 0 \text{ and } [D^i_Z(x)]_j < 0.$$

The case $x_j = 1$ and $[D^i_Z(x)]_j > 0$ follows symmetrically. By Lemma 27 we know that

$$\left(D^i_{Z/\{j\}}(x)\right)^\top \cdot \nabla V_j(x) < 0$$

Recall $V_j(x) = 0$ that and as a result for all $t \in [0, \delta]$ with $\delta > 0$ being sufficiently small $z_j(t) = 0$ and $V_j(z(t)) < 0$. As a result, $\delta > 0$ can be selected sufficiently small such as all the following hold,

1. $V_\ell\left(z(t)\right) = 0$ for all $\ell \in S$
2. $z_\ell(t) \in (0,1)$ for all $\ell \in S \cup \{i\}$.
3. ($z_\ell(t) = 0$ and $V_\ell(z(t)) < 0$) or ($z_\ell(t) = 1$ and $V_\ell(z(t)) > 0$) for all $\ell \in F$.
4. $z_\ell(t) = 0$ for all coordinates $\ell \geq i+1$.

■

Lemma 30 *Let the coordinates i, i' and $S \subseteq [\min(i,i')-1]$ and consider the direction $D^i_S(x), D^{i'}_S(x)$ of Definition 7 at a point $x \in [0,1]^n$. Then,*

$$\text{sign}\left(\left[D^i_S(x)\right]_i\right) = \text{sign}\left(\left[D^{i'}_S(x)\right]_{i'}\right)$$

Proof To simplify notation let $i < i'$ and $S = \{1, \ldots, i-1\}$ and let assume that S is even. Moreover let $D^i_S = (d_1, \ldots, d_{i-1}, d_i)$ and $D^{i'}_S = (d'_1, \ldots, d'_{i-1}, d'_{i'})$

By Definition 7 we know that

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_{i-1}} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_{i-1}} & d_{i-1} \\ \frac{\partial V_1(x)}{\partial x_i} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_i} & d_i \end{vmatrix} > 0 \tag{12}$$

Due to the fact that $(\nabla V_\ell(x))^\top \cdot (d_1, \ldots, d_{i-1}, d_i) = 0$ for all $\ell \in \{1, \ldots, i-1\}$,

$$\text{sign}\left(\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_{i-1}} & \ldots & \frac{\partial V_{i-1}(x)}{\partial x_{i-1}} & d_{i-1} \\ 0 & \ldots & 0 & d_1^2 + \ldots + d_{i-1}^2 + d_i^2 \end{vmatrix}\right) = \text{sign}(d_i) \tag{13}$$

implying that

$$\text{sign}\left(\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \cdots & \frac{\partial V_{i-1}(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_{i-1}} & \cdots & \frac{\partial V_{i-1}(x)}{\partial x_{i-1}} \end{vmatrix}\right) = \text{sign}(d_i) \tag{14}$$

Repeating the same argument for $(d'_1, \dots, d'_{i-1}, d'_{i'})$ we get that $\text{sign}(d_i) = \text{sign}(d'_{i'})$. ■

Lemma 31 *Let a pivot $x \in [0,1]^n$ with at least one unsatisfied coordinate. Let (i, S) the admissible pair of pivot x (Definition 16) 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 $D^i_S(x)^\top \nabla V_i(x) > 0$*

Proof Let $Z^\ell(x) = \{\text{coordinates } j \leq \ell - 1 \text{ such that } V_j(x) = 0\}$ and $F^\ell(x) = \{\text{coordinates } j \leq \ell - 1 \text{ that are satisfied at } x\}$. Since ℓ is minimum unsatisfied coordinate, by the definition of pivot (Definition 11) we known that $V_\ell(x) > 0$. As a result, $x_\ell \neq 1$ since otherwise coordinate i would be satisfied at x. In case $x_\ell \in (0,1)$ then the coordinate ℓ is not frozen and thus be Definition 16, the under examination coordinate at pivot x is $i = \ell \geq 1$. Thus the first item of Lemma 31 is established. Similarly in case $x_\ell = 0$ and $[D_{Z^\ell(x)}(x)]_\ell \geq 0$. As a result, we consider the case where $[D_{Z^\ell(x)}(x)]_\ell < 0$ and $x_\ell = 0$.

At first notice that in case $Z^\ell(x) = \varnothing$ then by Definition 7 we get that $[D^\ell(x)]_\ell = 1$ which contradicts with $[D_{Z^\ell(x)}(x)]_\ell < 0$. Also since $x_\ell = 0$, Assumption 2 implies that $x_j \in (0,1)$ for all coordinates $j \in Z^\ell(x)$.

Let assume that $x_{\ell-1} = 0$. Since $x_j \in (0,1)$ for all $j \in Z^\ell(x)$ coordinate $\ell - 1 \notin Z^\ell(x)$ and thus $Z^{\ell-1}(x) = Z^\ell(x)$. Lemma 30 together with the fact that $x_j \in (0,1)$ for $j \in Z^{\ell-1}$, imply 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

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

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 $x_i = 1$. Then $i \notin Z^{i+1}(x) = Z^\ell(x)$ since otherwise $x_i \in (0,1)$. As a result, $Z^i(x) = Z^{i+1}(x) = Z^\ell(x)$ and again by Lemma 30, $\text{sign}([D^i(x)]_i) = \text{sign}([D^\ell(x)]_\ell) = -1$. Thus coordinate i is not frozen and at the same time $V_i(x) > 0$ since coordinate i is satisfied at x and $V_i(x) \neq 0$ ($i \notin Z^\ell(x)$).

Now let us consider the case where $x_i \in (0, 1)$. Then 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 $D^\ell(x) := (d_1, \dots, d_i, d_\ell)$ and $D^i(x) := (\hat{d}_1, \dots, \hat{d}_i)$. Let us assume that $|Z^\ell(x)|$ is even (the case $|Z^\ell(x)|$ is even follows symmetrically). Then by Definition 7 we get that,

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_i(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_i(x)}{\partial x_i} & d_i \\ \frac{\partial V_1(x)}{\partial x_\ell} & \dots & \frac{\partial V_i(x)}{\partial x_\ell} & d_\ell \end{vmatrix} > 0 \tag{15}$$

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

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_i(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_i(x)}{\partial x_i} & d_i \\ 0 & \dots & 0 & d_1^2 + \dots + d_i^2 + d_\ell^2 \end{vmatrix} < 0 \tag{16}$$

implying that

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_i(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_i(x)}{\partial x_i} \end{vmatrix} < 0 \tag{17}$$

Since $|Z^\ell(x)|$ is even then $|Z^i(x)|$ is odd ($Z^\ell(x) = Z^i(x) \cup \{i\}$) and thus by Definition 7

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \hat{d}_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \hat{d}_i \end{vmatrix} < 0 \tag{18}$$

Multiplying Equation 18 with Equation 16 we get that,

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

■

Lemma 32 *Let a point $x \in [0, 1]^n$, a coordinate $i \in [n]$ and a set of coordinates $S \subseteq [i - 1]$ such that $V_j(x) = 0$ for all $j \in S \cup \{i\}$. Then for any coordinate $\ell > i$ such that $[D^\ell_{S \cup \{i\}}(x)]_\ell < 0$, $D^i_S(x)^\top \cdot V_\ell(x) < 0$.*

Proof To simplify notation let $D^\ell_{S\cup\{i\}}(x) = (d_1, \dots, d_i, d_\ell)$ and $D^i_S(x) = (\hat{d}_1, \dots, \hat{d}_{i-1}, \hat{d}_i)$. Let us assume that $|S \cup \{i\}|$ is even (the odd case follow symmetrically). By Definition 7 we get that,

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_i(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_i(x)}{\partial x_i} & d_i \\ \frac{\partial V_1(x)}{\partial x_\ell} & \dots & \frac{\partial V_i(x)}{\partial x_\ell} & d_\ell \end{vmatrix} > 0 \tag{19}$$

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

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_i(x)}{\partial x_1} & d_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_i(x)}{\partial x_i} & d_i \\ 0 & \dots & 0 & d_1^2 + \ldots + d_i^2 + d_\ell^2 \end{vmatrix} < 0 \tag{20}$$

implying that

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_i(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_i(x)}{\partial x_i} \end{vmatrix} < 0 \tag{21}$$

Since $S \cup \{i\}$ is even then $|S|$ is odd and thus by Definition 7

$$\begin{vmatrix} \frac{\partial V_1(x)}{\partial x_1} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_1} & \hat{d}_1 \\ \vdots & & \vdots & \vdots \\ \frac{\partial V_1(x)}{\partial x_i} & \dots & \frac{\partial V_{i-1}(x)}{\partial x_i} & \hat{d}_i \end{vmatrix} < 0 \tag{22}$$

Multiplying Equation 22 with Equation 20 we get that,

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

■

Appendix I. Proof of Lemma 17

Proof [Proof of Lemma 17] Lemma 17 follows by Lemma 28. By Definition 11 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 28 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 28 the number of such z is at most $2^n/\text{Vol}^n\left(\frac{2\sigma_{min}^2}{\sqrt{n}L\sigma_{max}^2}\right)$). Since they are only 2^n choices for the boundary coordinates, the overall number of pivots is at most $4^n/\text{Vol}^n\left(\frac{2\sigma_{min}^2}{\sqrt{n}L\sigma_{max}^2}\right)$. ■

Appendix J. Proof of Lemma 19

We first present some auxiliary lemmas.

Lemma 33 *Let a set of coordinates S, a coordinate i and a point $x \in [0, 1]^n$ such that $x_j \in (0, 1)$ for all $j \in S \cup \{i\}$, $V_j(x) = 0$ for all $j \in S$ and $x_j \in \{0, 1\}$ for all $j \notin S \cup \{i\}$. Then there exists* *a $t^* \in (0, C)$, where C is constant depending on the parameters $\sigma_{min}, \sigma_{max}$ and L, such that for $t \in [0, t^*]$ the initial value problem*

$$\dot{\gamma}(t) = D_S^i(\gamma(t)) \quad \textit{with} \quad \gamma(0) = x$$

admits a unique solution. Moreover the following hold,

1. $\gamma_j(t) \in (0, 1)$ *for all* $j \in S \cup \{i\}$ *and* $t \in ([0, t^*])$.
2. $\gamma_j(t^*) = 0$ *or* 1 *for some* $j \in S \cup \{i\}$

Proof [Proof of Lemma 19] Given the pivot $x \in [0, 1]^n$ with at least one unsatisfied variable and let (i, S) denote its admissible pair. By Lemma 31 we know that the under examination variable i admits $i \geq 1$. By Lemma 29 there exists a unique trajectory $\dot{z}(t) = D_S^i(z(t))$ with $z(0) = x$ for $t \in [0, t^*]$ such that for all $t \in (0, t^*]$,

1. $V_j(z(t)) = 0$ for all coordinates $j \in S$.
2. $z_j(t) \in (0, 1)$ for all coordinates $j \in S \cup \{i\}$.
3. $(z_j(t) = 0$ and $V_j(z(t)) < 0)$ or $(z_j(t) = 1$ and $V_j(z(t)) > 0)$ for all $j \in [i - 1]/S$.
4. $z_j(t) \in \{0, 1\}$ for all coordinates $j \notin S \cup \{i\}$.
5. $z_j(t) = 0$ for all coordinates $j \geq i + 1$.

Since $z_j(t^*) \in (0, 1)$ for all $j \in S \cup \{i\}$, Lemma 33 establishes that there exists a unique trajectory $\dot{z}(t) = D_S^i(z(t))$ with $z(t^*) = x$ for $t \in [t^*, \hat{t}]$ such that

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

By Item 3 we know that $(z_j(t^*) = 0$ and $V_j(z(t^*)) < 0)$ or $(z_j(t^*) = 1$ and $V_j(z(t^*)) > 0)$ for all coordinates $j \in \{\ell \leq i - 1\}/S$.

Let $t' \in (t^*, \hat{t})$ be the first time such that $V_j(z(t')) = 0$ for some $j \in \{\ell \leq i - 1\}/S$. Since $z_j(t^*) \in \{0, 1\}$ and $j \notin S$ we get that $z_j(t') = z_j(t^*) \in \{0, 1\}$. Applying Assumption 2 in $z(t')$ we get that there exists a unique $j \in \{\ell \leq i - 1\}/S$ such that $V_j(z(t')) = 0$. As a result, $\{\ell \leq i - 1 \ : \ V_\ell(z(t')) = 0\} = S \cup \{j\}$ and by Definition 11 we get that $z(t')$ is a pivot.

In case there is no such t^*, then $V_j(z(\hat{t})) < 0$ for all $j \in \{\ell \leq i - 1\}/S$ with $z_j(\hat{t}) = z_j(t^*) = 0$ and $V_j(z(\hat{t})) > 0$ for all $j \in \{\ell \leq i - 1\}/S$ with $z_j(\hat{t}) = z_j(t^*) = 1$. As a result, the set $\{V_\ell(x) = 0 \ : \ \ell \leq i - 1\} = S$ and since $z_j(\hat{t}) \in \{0, 1\}$ for some coordinate $j \in S \cup \{i\}$, Definition 11 implies that $z(\hat{t})$ is a pivot. ■

J.1. Proof of Lemma 33

Consider the ball $\mathcal{B} := \mathcal{B}(x, \rho)$ with $\rho = \sigma_{\min}/2Ln$. Since $V_j(x) = 0$ for all $j \in S$, by Assumption 1 we know that the matrix $J_S(x)$ admits singular values greater than $\sigma_{\min}$. Moreover by Lemma 24 for all $x' \in \mathcal{B}$ the matrix $J_S(x')$ admits singular values greater than $\sigma_{\min}/2$. Thus by Lemma 25, for $\hat{x}, x' \in \mathcal{B}$,

$$D_S^i(\hat{x}) - D_S^i(x') \leq M \cdot \|\hat{x} - x'\|_2$$

where $M := \Theta\left(\frac{\sigma_{\max}}{\sigma_{\min}^2} \cdot n \cdot L\right)$. Since $D_S^i(\cdot)$ is M-Lipschitz continuous in $\mathcal{B}$ and $\|D_S^i(\cdot)\|_2 = 1$, the Picard–Lindelöf theorem implies that for all $t \in [0, 1/M]$ the initial value problem

$$\dot{\gamma}(t) = D_S^i(\gamma(t)) \text{ and } \gamma(0) = x$$

admits a unique solution.

In case there exists $t^* \in [0, 1/M]$ such that $\gamma_j(t^*) \in 0, 1$ for some coordinate $j \in S \cup \{i\}$ then we are done. Otherwise let $t^* = 1/M$. Since $\dot{\gamma}(t) = D_S^i(\gamma(t))$ it ensured that $V_j(\gamma(t^*)) = 0$ for all $j \in S$. Let assume that we repeat the same $k := \frac{2 \cdot 2^n}{M^2 \mathrm{Vol}^n(\rho)}$ times where $\rho = \Theta\left(\frac{\sigma_{\min}^3}{\sqrt{n}\sigma_{\max}^2 L}\right)$ times until $\gamma_j(t^*) \in \{0, 1\}$ for some coordinate $j \in S \cup \{i\}$. In case the *exit condition* was satisfied at some iteration then we are done. Let assume that the latter never happen which means that

$$\dot{\gamma}(t) = D_S^i(\gamma(t)) \text{ and } \gamma(0) = x$$

for $t \in [0, T]$ where $T := \frac{2^n}{M^2 \mathrm{Vol}^n(\rho)}$ and $\rho = \Theta\left(\frac{\sigma_{\min}^3}{\sqrt{n}\sigma_{\max}^2 L}\right)$. Up next we show that the latter leads to contradiction.

To simplify notation let $S := (1, \ldots, i-1)$ and let $D_S^i(x)$ be denoted as $D(x)$. Let also M denote the Lipschitz constant of $D(\cdot)$ established in Lemma 25. To further simplify notation we denote as $\Phi_\ell(x)$ the gradient of $V_\ell(x)$ with respect to the coordinates $\{1, \ldots, i\}$,

$$\Phi_\ell(x) := \left(\frac{\partial V_\ell(x)}{\partial x_1}, \ldots, \frac{\partial V_\ell(x)}{\partial x_i}\right)$$

Since $V_j(\gamma(t)) = 0$ for all coordinates $j \in S$, Assumption 1 ensures that at any point $x \in [0, 1]^i$ the matrix

$$\Phi(x) := \begin{pmatrix} \Phi_1(x) \\ \Phi_2(x) \\ \vdots \\ \Phi_{i-1}(x) \end{pmatrix} := \begin{pmatrix} \nabla V_1(x) \\ \nabla V_2(x) \\ \vdots \\ \nabla V_{i-1}(x) \end{pmatrix} \tag{23}$$

admits singular values greater than $\sigma_{\min}$ and smaller than $\sigma_{\max}$. 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. ■

Lemma 34 *Let $\delta \leq 1/4$ and $p \in [0,1]^n$ such that $\|\gamma(t_0) - p\| \leq \frac{\delta}{2M}$. Then there exists $t^* \in [-1/M, 1/M]$ such that*

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

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 $t' = \inf_{0 \leq t \leq 1/M}\{\|\gamma(t + t_0) - \gamma(t_0)\|_2 \geq \frac{1}{4M}\}$. By the triangle inequality,

$$\|\gamma(t' + t_0) - p\|_2 \geq \|\gamma(t' + t_0) - \gamma(t_0)\|_2 - \|\gamma(t_0) - p\|_2 \geq \frac{1}{4M} - \frac{\delta}{2M} \geq \frac{\delta}{2M}$$

and thus there exists $\hat{t}_1 \in [0, t']$ such that

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

By expanding on the first item we get,

$$\begin{aligned}\|\gamma(t+t_0)-p\|^2 &= \|\gamma(t+t_0)-\gamma(\hat{t}_1+t_0)+\gamma(\hat{t}_1+t_0)-p\|^2\\ &= \|\gamma(t+t_0)-\gamma(\hat{t}_1+t_0)\|^2+\|\gamma(\hat{t}_1+t_0)-p\|^2\\ &+ 2\left(\gamma(t+t_0)-\gamma(\hat{t}_1+t_0)\right)^\top\cdot\left(\gamma(\hat{t}_1+t_0)-p\right)\\ &= \|\gamma(t+t_0)-\gamma(\hat{t}_1+t_0)\|^2+\frac{\delta^2}{4M^2}\\ &+ 2\left(\gamma(t+t_0)-\gamma(\hat{t}_1+t_0)\right)^\top\cdot\left(\gamma(\hat{t}_1+t_0)-p\right)\end{aligned}$$

Since $\|\gamma(t+t_0)-p\|^2 \le \frac{\delta^2}{4M^2}$ the latter implies that

$$\frac{1}{t_1-t}\left(\int_t^{t_1} D(\gamma(t_0+s))\partial s\right)^\top\cdot(\gamma(t_1+t_0)-p) \ge \frac{1}{t_1-t}\|\gamma(t+t_0)-\gamma(t_1+t_0)\|^2$$

and thus by taking $t \to \hat{t}_1$ we get that

$$D^\top\left(\gamma(t_0+\hat{t}_1)\right)\cdot\left(\gamma(t_0+\hat{t}_1)-\gamma\right) \ge 0$$

As a result, we have shown that there exists $\hat{t}_1 \in [0, 1/M]$ such that

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

where the first item comes from the fact that $\|\gamma(t+t_0)-p)\| \le \delta/2M$ for all $t \le \hat{t}_1$ and the fact that $\|\gamma(t_0)-p\| \le \delta/2M$.

Symmetrically we can prove that there exists $\hat{t_2} \in [0, 1/M]$ such that

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

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 $t \in [-\hat{t}_2, \hat{t}_1]$. ■

Up next we present the main lemma of the section.

Lemma 35 *Let $\rho = \Theta\left(\frac{\sigma_{min}^3}{\sqrt{n}\cdot\sigma_{max}^2\cdot L}\right)$ and a point $p \in \mathbb{B}(\gamma(t_0), \rho/2)$ with $p \notin \gamma[t_0-1/M, t_0+1/M]$. Then there is a coordinate $\ell \le i-1$ such that $V_\ell(p) \neq 0$.*

Proof Let $\delta = M\cdot\rho$ where M is the Lipschitz constant of $D(x)$. Let us assume that $V_\ell(p) = 0$ for all coordinates $j \le i-1$. By Lemma 34 there exists $t^* \in [t_0-1/M, t_0+1/M]$ such that

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

Since $p \notin \gamma[t_0 - 1/M, t_0 + 1/M]$ we know that $(\gamma(t^* + t_0) - p) \neq 0$. Combining the fact that $D^\top(\gamma(t^* + t_0)) \cdot (\gamma(t^* + t_0) - p) = 0$ with $D^\top(\gamma(t^* + t_0)) \cdot \Phi_\ell(\gamma(t^* + t_0)) = 0$ for all $\ell \in [i-1]$ and the fact that $\Phi_1(\gamma(t_0 + t^*)), \dots, \Phi_{i-1}(\gamma(t_0 + t^*))$ are linearly independent we get that,

$$p - \gamma(t_0 + t^*) = \sum_{j=1}^{i-1} \mu_j \cdot \Phi_j \left(\gamma(t_0 + t^*)\right)$$

with $\|\mu\|_2 \neq 0$.

Then due to the fact that $\|\gamma(t_0) - p\|_2 \leq \rho = \frac{\delta}{M}$ and $\|\gamma(t_0 + t^*) - \gamma(t_0)\|_2 \leq \frac{\delta}{M}$ we get,

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

As a result,

$$\|\mu\|_2 \leq \frac{2\delta}{\sigma_{\text{min}} \cdot M} \tag{24}$$

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

$$\left| V_\ell(p) - V_\ell\left(\gamma(t_0 + t^*)\right) - \Phi_\ell^\top\left(\gamma(t_0 + t^*)\right) \cdot \sum_{j=1}^{i-1} \mu_j \Phi_j\left(\gamma(t_0 + t^*)\right) \right| \leq \Theta\left(L \cdot \|\gamma(t_0 + t^*) - p\|_2^2\right)$$

Since $V_\ell(p) = V_\ell\left(\gamma(t_0 + t^*)\right) = 0$

$$\left| \Phi_\ell^\top\left(\gamma(t_0 + t^*)\right) \cdot \sum_{j=1}^{i-1} \mu_j \Phi_j\left(\gamma(t_0 + t^*)\right) \right| \leq \Theta\left(L \cdot \|\sum_{j=1}^{i-1} \mu_j \cdot \Phi_j\left(\gamma(t_0 + t^*)\right)\|^2 \right) \leq \Theta\left(L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|_2^2\right)$$

meaning that $\left|\left[\Phi(\gamma(t_0 + t^*))\Phi^\top(\gamma(t_0 + t^*)) \cdot \mu\right]_\ell\right| \leq \Theta\left(L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|_2^2\right)$ and thus

$$\sigma_{\text{min}}^2 \|\mu\|_2 \leq \|\Phi(\gamma(t_0 + t^*))\Phi^\top(\gamma(t_0 + t^*)) \cdot \mu\|_2 \leq \Theta\left(\sqrt{n} \cdot L \cdot \sigma_{\text{max}}^2 \cdot \|\mu\|_2^2\right) \rightarrow \|\mu\|_2 \geq \Theta\left(\frac{\sigma_{\text{min}}^2}{\sqrt{n} \cdot L \cdot \sigma_{\text{max}}^2}\right)$$

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 with final step in the proof of Lemma 33. 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 $t_k := t_1 + \frac{2k}{M}$ and let the ball $\mathbb{B}_k := \mathbb{B}(\gamma(t_k), \rho/2)$ 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 $\mathbb{B}_k \cap \mathbb{B}_{k'} = \varnothing$ for $k \neq k'$ since otherwise by Lemma 35 $V_\ell(\gamma(\tau)) \neq 0$ for some coordinate $\ell \in S$ and $\tau > 0$. However the latter is a contradiction due to the fact that $\mathbb{B}_i \cap [0,1]^i$ are disjoint sets with volume greater than $\frac{\text{Vol}^n(\rho/2)}{2^n}$.

Appendix K. Proof of Lemma 21

Lemma 36 *Any pivot $x \in [0,1]^n$ admits in-degree at most 1. In other words, in case $p = \text{Next}(x_1)$ and $p = \text{Next}(x_2)$ for some pivots x_1, x_2 then $x_1 = x_2$.*

Proof Let the pivot $p = \text{Next}(x_1) = \text{Next}(x_2)$ while $x_1 \neq x_2$. Moreover let (i, S) the admissible pair of pivot x_1 and (i', S') the admissible pair of pivot x_2. By Definition 18 there is a trajectory $\dot{z}(t) = D_S^i(z(t))$ with $z(0) = x_1$ and $\dot{y}(t) = D_{S'}^{i'}(y(t))$ with $y(0) = x_2$ while $z(t_1) = y(t_2) = p$ for some $t_1, t_2 > 0$. Up next we show that the latter always leads to contradiction.

Let $S = S'$ and $i = i'$. As a result, both of the trajectories $z(t), y(t)$ admit $\dot{z}(t) = D_S^i(z(t))$ and $\dot{y}(t) = D_S^i(y(t))$. Also recall that $z(t_1) = y(t_2) = p$. Without loss of generality let $t_1 < t_2$.

Let also $F(x) := -D_S^i(x)$. By Lemma 25 and the Picard–Lindelöf theorem, there exists a unique function $w(t)$ such that $w(t) := -D_S^i(w(t))$ and $w(0) = p$. Since $w(t_1) = x_1, w(t_2) = x_2$ and $t_1 < t_2$ we get that $y(t_2 - t_1) = x_1$. This is a contradiction since by Lemma 19, $y(t)$ is not a pivot for $t \in (0, t_2)$.

Let $M := (S'/S) \cup (S/S')$, up next we show that both cases $M \neq \varnothing$ and ($M = \varnothing$ and $i \neq i'$) lead to contradiction.

- $|M| \geq 2$:
  - $i = i'$: We first show that there exist two coordinates $\ell_1, \ell_2 \in M$ with $\ell_1 \neq \ell_2$ such that

    1. $V_{\ell_1}(p) = V_{\ell_2}(p) = 0$
    2. $p_{\ell_1} \in \{0, 1\}$
    3. $p_{\ell_2} \in \{0, 1\}$

    Without loss of generality let $\ell_1 \in S'/S$ and thus $\ell_1 \neq i$. Since $p = \text{Next}(x_2)$ and $\ell_1 \in S'$ then by Lemma 19, $V_{\ell_1}(p) = 0$. At the same time since $\ell_1 \notin S$ and $\ell_1 \neq i$, by Lemma 19 we get that $p_{\ell_1} \in \{0, 1\}$. Since $|M| \geq 2$ we can exclude ℓ_1 form M and repeat the same argument for ℓ_2.

    Now consider the set of coordinates $A := S \cup S' \cup \{i\}$. By Lemma 19 all coordinates $j \notin A$ admit $p_j \in \{0, 1\}$. Moreover by Lemma 19 all coordinates $\ell \in A/\{i\}$ admit $V_\ell(p) = 0$. Then by Assumption 2 there exists a unique coordinate $\ell \in A$ such that $p_\ell \in \{0, 1\}$ which is a contradiction since $\ell_1, \ell_2 \in A$ and $\ell_1 \neq \ell_2$.

  - $i' > i$: Consider the set $A := \{j \leq i' - 1 : \ V_j(p) = 0\}$. Lemma 19 ensures that $S' \subseteq A$ and that $p_j \in \{0, 1\}$ for all $j \notin S'$. Thus, $p_j \in \{0, 1\}$ for all $j \notin A$. Since $i' > i$ and $p = \text{Next}(x_1)$ by Lemma 19 we know that $p_{i'} = 0$. By applying Assumption 2 on $A \cup \{i'\}$, we get that $p_j \in (0, 1)$ for all coordinates $j \in A$. As a result,

    1. $S \subseteq A$: Let $j \in S$ then $j \leq i - 1$ and by Lemma 19 $V_j(p) = 0$. Thus, $j \in A$.
    2. $A \subseteq S$: Let $j \in A$ and $j \notin S$. Since $j \in A$ we have already proven that $p_j \in (0, 1)$. At the same time since (i, S) is an admissible pair for pivot x_1, Lemma 19 ensures that $p_j \in \{0, 1\}$ for all $j \notin S$. The latter is a contradiction.

3. $A \subseteq S'$: Let $j \in A$ and $j \notin S'$. Since $j \in A$ we have already proven that $p_j \in (0,1)$. At the same time since (i', S') is an admissible pair for pivot x_2, Lemma 19 ensures that $p_j \in \{0,1\}$ for all $j \notin S'$. The latter is a contradiction.
4. $S' \subseteq A \cup \{i\}$: Let $j \in S'$ and $j \notin A \cup \{i\}$. Since $j \in S'$ by Lemma 19 we get that $V_j(p) = 0$. Since $j \notin A \cup \{i\}$ we have already proven above that $p_j \in \{0,1\}$. Since $i < i'$ by Lemma 19 coordinate i is satisfied at p. In case $p_i \in \{0,1\}$ then Assumption 2 is violated for $A \cup \{j, i\}$ since $p_i, p_j \in \{0,1\}$. In case $V_i(p) = 0$ then Assumption 2 is violated for the set $A \cup \{j, i, i'\}$ since $p_{i'}, p_j \in \{0,1\}$.

All the above imply that $S \subseteq S' \subseteq S \cup \{i\}$ which contradicts with $|M| \geq 2$.

- $|S'/S| = 1$ and $i' > i$: Consider $\ell \in S'/S$. Since $\ell \in S'$, by Lemma 19 we get that $V_\ell(p) = 0$. At the same time since i is the under examination coordinate at x_1 and $i' > i$ by Lemma 19 we get that $p_{i'} = 0$.

  - $\ell \neq i$: Since $\ell \neq i$ and $\ell \notin S$ by Lemma 19 we get that $p_\ell \in \{0,1\}$. Since $p = \text{Next}(x_1)$ and i is the under examination variable at x_1, by Lemma 19 we know that $p_j = 0$ for all $j \geq i+1$. As a result, $p^*_{i'} = 0$. Now consider the set of coordinates $A = \{j \leq i' - 1 : \; V_j(p) = 0\} \cup \{i'\}$. By Lemma 19 we get that $p_j \in \{0,1\}$ for all coordinates $j \notin A$. Since both $p_{i'}, p_\ell \in \{0,1\}$, the latter contradicts with Assumption 2 applied on set A.

  - $\ell = i$: In this case $S' = S \cup \{i\}$. Since $i' > i$ and i is the under examination coordinate at x_1 by Lemma 19 we get that $p_{i'} = 0$. Again by Lemma 19 we know that $y_i(t) \in (0,1)$ for all $t \in (0, t_2)$ and thus $[D^{i'}_{S'}(p)]_{i'} = [D^{i'}_{S \cup \{i\}}(p)]_{i'} \leq 0$. Then Assumption 3 implies $[D^{i'}_{S'}(p)]_{i'} = [D^{i'}_{S \cup \{i\}}(p)]_{i'} < 0$. Then Lemma 32 implies that $D^i_S(p)^\top \cdot \nabla V_i(p) > 0$.

    Since $i \in S'$, Lemma 19 implies that $V_i(p) = 0$. Since i is the under examination coordinate at x_1, by Lemma 31 we get that $V_i(z(t)) > 0$ for all $t \in (0, \delta)$ where δ is sufficiently small. Since $V_i(p) = 0$ and by Lemma 19 $z_i(t) \in (0,1)$ for all $t \in (0, t_1)$, we get that $D^i_S(p)^\top \nabla V_i(p) \leq 0$ which is a contradiction.

- $|S/S'| = 1$ and $i' > i$: Let $S = S' \cup \{\ell\}$. Lemma 19 implies that $V_j(p) = 0$ for all $j \in S'$ and $V_\ell(p) = 0$ since $\ell \in S$. At the same time since $i' > i$, Lemma 19 implies that $p_{i'} = 0$ and $V_\ell(p) = 0$ since $\ell \notin S' \cup \{i'\}$. Consider the set $A := S' \cup \{i'\} \cup \{j\}$ then $p_j \in \{0,1\}$ for all $j \notin A$ and $V_j(p) = 0$ for $j \in S' \cup \{\ell\}$. However the fact that $p_{i'}, p_\ell \in \{0,1\}$ contradicts Assumption 2.

- $|M| = 1$ and $i' = i$: Without loss of generality let $\ell \in S'/S$. Since $p = \text{Next}(x_2)$ by Lemma 19 we get that $V_\ell(p) = 0$ since $\ell \in S'$. Since $\ell \notin S$, $i \neq \ell$ and $p = \text{Next}(x_1)$, Lemma 19 implies that $p_\ell \in \{0,1\}$.

  Let us consider the following mutually exclusive cases,

  - $p_i \in \{0,1\}$: Consider the set of coordinates $A = \{j \leq i - 1 : \; V_j(p) = 0\} \cup \{i\}$. Lemma 19 guarantees that $V_j(p) = 0$ for all coordinates $j \in S$ and $p_j \in \{0,1\}$ for all $j \notin S \cup \{i\}$. Since $S \cup \{i\} \subseteq A$, all coordinates $j \notin A$ admit $p_j \in \{0,1\}$. Since both $i, \ell \in A$ and $p_i, p_\ell \in \{0,1\}$, the latter contradicts Assumption 2.

  - $p_i \in (0,1)$ and $V_i(p) = 0$: Consider the set $A := \{j \leq i-1 : \ V_j(p) = 0\} \cup \{i\} \cup \{i+1\}$. By Lemma 19, we get that $V_j(p) = 0$ for all coordinates $j \in S$ and $p_j \in \{0,1\}$ for all $j \notin S$. Since $S \subseteq A$, all coordinates $j \notin A$ admit $p_j \in \{0,1\}$. At the same time by Lemma 19 we get that $p_{i+1} = 0$. Since both $i+1, \ell \in A$ and $p_{i+1}, p_\ell \in \{0,1\}$, the latter contradicts Assumption 2. Note that $\ell \neq i+1$ comes from the fact that (i, S') is the admissible pair at x_2 while $\ell \in S'$. Thus, $\ell \leq i-1$.

  - $x_i \in (0,1)$ and $V_i(x^*) > 0$: Without loss of generality we assume that $p_\ell = 0$ (symmetrically if $p_\ell = 1$). Since $\ell \notin S \cup \{i\}$, $[D^i_S(z(t))]_\ell = 0$ (Definition 7) and thus $z_\ell(t) = 0$ for all $t \in [0, t_1]$. By Lemma 19 we know that ℓ remains satisfied during the trajectory $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x_1$ meaning that $V_\ell(z(t)) \leq 0$. Since $V_\ell(p) = 0$ we get that $D^i_S(p)^\top \cdot \nabla V_\ell(p) \geq 0$.

    Since $\ell \in S'$ by Lemma 19 we get that $y_\ell(t) \in (0,1)$ for $t \in (0, t_2)$. Since $p_\ell = y_\ell(t_2) = 0$ and $S' = S \cup \{\ell\}$, $[D^i_{S'}(p)]_\ell = [D^i_{S \cup \{\ell\}}(p)]_\ell \leq 0$. Since $p_\ell = 0$, by Assumption 3 we additionally get that $[D^i_{S'}(p)]_\ell = [D^i_{S \cup \{\ell\}}(p)]_\ell < 0$. Then by applying Lemma 27 for $Z := S \cup \{\ell\}$ we get that $D^i_S(p)^\top \cdot \nabla V_\ell(p) < 0$ which is a contradiction.

- $|M| = 0$ and $i' > i$: Since i is the under examination variable at x_1, Lemma 19 implies that $p_{i'} = 0$. Lemma 19 also ensures that $y_{i'}(t) \in (0,1)$ for all $t \in (0, t_2)$. Thus we additionally get that $[D^{i'}_S(p)]_{i'} \leq 0$ with combined with Assumption 3 implies $[D^{i'}_S(p)]_{i'} < 0$. Since all coordinates $j \in S$ are less than $i-1$, by Lemma 30 we get that $\text{sign}([D^i_S(p)]_i) = \text{sign}([D^{i'}_S(p)]_{i'})$ and thus $[D^i_S(p)]_i < 0$.

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

  - $p_i \in (0,1)$ and $V_i(p) = 0$.
  - $p_i = 1$ and $V_i(p) \geq 0$.
  - $p_i = 0$ and $V_i(p) \leq 0$.

  Since $i \notin S$ Lemma 19 implies that $p_i \in \{0,1\}$ which excludes the first case. In case $p_i = 1$ by Lemma 19 we get that $[D^i_S(p)]_i \geq 0$ ($z_i(t) \in (0,1)$ for all $t \in (0, t_1)$) which contradicts with $[D^i_S(p)]_i < 0$. Up next we exclude the third case where $p_i = 0$ and $V_i(p) \leq 0$.

  Since i is the under examination coordinate at point x_1 and $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x_1$, by Lemma 31 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 $V_i(p) \leq 0$ implies that $V_i(p) = 0$ (otherwise $z(t)$ is a pivot for some $t \in (0, t_1)$). Now consider the set $A := S \cup \{i\} \cup \{i'\}$ and notice that $p_j = 0$ for all $j \notin A$. The fact that $p_i = p_{i'} = 0$ and $V_j(p) = 0$ for all $j \in S \cup \{i\}$ contradicts with Assumption 2.

■

Appendix L. Proof of Lemma 22

Let $(0,\dots,0) = \mathrm{Next}(x)$ for some pivot $x \in [0,1]^n$ and let (i,S) its admissible pair. As a result, $\dot{z}(t) = D^i_S(z(t))$ with $z(0) = x$ and $z(t^*) = (0,\dots,0)$ for some $t^* > 0$.

Let $|S| = 0$ then $[D^i_\varnothing(z(t))]_i = 1$. By Lemma 19, $z_i(t) \in (0,1)$ for all $t \in (0,t^*)$ and since $z_i(t^*) = 0$ we get that $[D^i_\varnothing(z(t))]_i \le 0$.

Let $|S| \ge 1$. By Lemma 19, $V_j(z(t^*)) = 0$ for all coordinates $j \in S$. Thus $V_j(0,\dots,0) = 0$ for all coordinates $j \in S$. Let the set $A = \{j \le i-1 : \; V_j(0,\dots,0) = 0\} \cup \{i\}$. Since $S \ge 1$ then $|A| \ge 2$. Then Assumption 2 is violated for $(0,\dots,0)$.

Appendix M. 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 37 ((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., $\|V_i(x')\| \le \varepsilon$, or $x'_i = 0$ and $V_i(x') < \varepsilon$, or $x'_i = 1$ and $V_i(x') > -\varepsilon$.*
- ***(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 38 *We say that a point x is an α-approximate solution of* $\mathrm{VI}(V,[0,1]^n)$ *if and only if $V(x)^\top(x-y) \le \alpha$.*

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 39.

Theorem 39 *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.*

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 $(\varepsilon,\gamma)$-exit point as per Definition 37 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 37 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 37 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 37 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 37 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$x^{(m)}$
```

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 VI(V, $[0,1]^n$) (see Theorem 9) 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 VI(V, $[0,1]^n$).

The proof of Theorem 39 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 40 (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 $\|w - w'\|_2 \leq \zeta$. Then, for every $\eta > \zeta$ and every $T > 0$, there exists a step size $\gamma > 0$ such that*

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

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

Given that $D_S^i(x)$ is Lipschitz (see Lemma 25) we can apply Theorem 40 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 9 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 40 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 9 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 40 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 VI$(V, [0,1]^n)$ we have that $x^{(M)}$ is an α-approximate solution to 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 9 divided by $\gamma = \gamma_m$.

■