Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions

Ioannis Panageas[1] and Georgios Piliouras[2]

1 **MIT, USA & Singapore University of Technology and Design, Singapore**
panageasj@gmail.com

2 **Singapore University of Technology and Design, Singapore**
georgios.piliouras@gmail.com

Abstract

Given a twice continuously differentiable cost function f, we prove that the set of initial conditions so that gradient descent converges to saddle points where $\nabla^2 f$ has at least one strictly negative eigenvalue, has (Lebesgue) measure zero, even for cost functions f with non-isolated critical points, answering an open question in [12]. Moreover, this result extends to forward-invariant convex subspaces, allowing for weak (non-globally Lipschitz) smoothness assumptions. Finally, we produce an upper bound on the allowable step-size.

1998 ACM Subject Classification G.1.6 [Numerical Analysis]: Optimization–Gradient methods

Keywords and phrases Gradient Descent, Center-stable manifold, Saddle points, Hessian

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.2

1 Introduction

The interplay between the structure of saddle points and the performance of gradient descent dynamics is a critical and not well understood aspect of non-convex optimization. Despite our incomplete theoretical understanding, in practice, the intuitive nature of the gradient descent method (and more generally gradient-like algorithms[1]) make it a basic tool for attacking non-convex optimization problems for which we have very little understanding of the geometry of their saddle points. In fact, these techniques become particularly useful as the equilibrium structure becomes increasingly complicated, e.g., such as in the cases of nonnegative matrix factorization [11] or congestion/potential games [25], where symmetries in the nature of non-convex optimization problems give rise to continuums of saddle points with complex geometry. In these cases, especially, the simplistic, greedy attitude of the gradient descent method, which is by design agnostic towards the global geometry of the cost function minimized, comes rather handy. As we move forward in time, the cost keeps decreasing and convergence is guaranteed.

This simplicity, however, comes at least seemingly at a significant cost. For example, it is well known that there exist instances where bad initialization of gradient descent converges to saddle points [18]. Despite the existence of such worst case instances in theory, practitioners have been rather successful at applying these techniques across a wide variety of problems [23]. Recently, Lee et. al. [12] have given a rather insightful interpretation of the effectiveness of gradient descent methods in terms of circumventing the saddle equilibrium problem using

[1] A gradient-like system is a system where for each non-equilibrium initial condition the dynamic will move towards a new state whose cost is strictly less than that of the initial state.

8th Innovations in Theoretical Computer Science Conference (ITCS 2017).
Editor: Christos H. Papadimitrou; Article No. 2; pp. 2:1–2:12
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

tools from topology of dynamical systems. At a glance, the paper argues the following intuitively clear message: *The instability of (locally unstable) saddle points translates to a global phenomenon and the probability of converging to such a saddle point given a randomly chosen (random not over a local neighborhood but over the whole state space) initial condition is zero.*

This message is clear, concise, and satisfying in the sense that it transcribes the practical success of the gradient descent method to a concrete theoretical guarantee. As is usually the case, such high level statements come with an asterisk of necessary technical conditions on the cost function f minimized.

Formally, a cost function f is said to satisfy the "strict saddle" property if each critical point[2] x of f is either a local minimizer, or a strict saddle, i.e., $\nabla^2 f(x)$ has at least one strictly negative eigenvalue. In this case, Lee et. al. argue that if $f : \mathbb{R}^d \to \mathbb{R}$ is a twice continuously differentiable function then gradient descent with constant step-size α (defined by $x_{k+1} = x_k - \alpha \nabla f(x_k)$) with a random initialization and sufficiently small constant step-size converges to a local minimizer or negative infinity almost surely.

Critically, for this result to apply, f is required to have isolated saddle points, ∇f is assumed to be globally L-Lipschitz[3] and the step-size α is taken to be less than $1/L$. These regularity conditions soften somewhat the impact of the statement both theoretically as well as in practice. First, although the assumption of isolated fixed points is indeed generic for abstract classes of cost functions, in several special cases of practical interest where the cost function has some degree of symmetry (e.g., due to scaling invariance) this assumption is not satisfied. For this reason, the question of whether the assumption of isolated equilibria is indeed necessary was explicitly raised in [12]. Moreover, the assumption of global Lipschitz continuity for ∇f is not satisfied even by low degree polynomials (e.g., cubic). Finally, a natural question is how tight is the assumption on the step-size?

In this work we provide answers to all the above questions. We show that the assumption of isolated saddle points is indeed not necessary to argue generic convergence to local minima. To argue this, we need to combine tools from dynamical systems, topology, analysis and optimization theory. Moreover, we show that the globally Lipschitz assumption can be circumvented as long as the domain is convex and forward invariant with respect to gradient descent. This proposition makes our results readily applicable to many standard settings. This extension holds even for non-coercive functions.[4] Finally, using linear algebra and eigenvalue analysis we provide an upper bound on the allowable step-size (for these results to hold).

It is important to clearly state that the focus of this work is more theoretical than applied. Specifically, although GD is guaranteed to converge to local minima this convergence can in the worst case (worst case initial conditions, instances) take exponential time as finding local minima of non-convex functions is an NP-hard problem. Nevertheless, we believe that this combination of topological techniques (standard in continuous-time dynamics) and discrete optimization techniques, besides solving an interesting open question, will hopefully be of use for other algorithmic problems as well. Deep learning, due to the singularity of the Hessian of the loss function in practice [24], is a domain where such techniques could lead to new insights and practical algorithms.

[2] x is a critical point of f if $\nabla f(x) = 0$.

[3] That is, f satisfies $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2$.

[4] As pointed out by a referee this extension of Lee et al. is relatively easy when a function is coercive, since the sublevel sets are compact.

1.1 Related work

First-order descent methods can indeed escape strict saddle points when assisted by near isotropic noise. [21] establishes convergence of the Robbins-Monro stochastic approximation to local minimizers for strict saddle functions, whereas [10] establishes convergence to local minima for perturbed versions of multiplicative weights algorithm in generic potential games. Recently, [7] quantified the convergence rate of perturbed stochastic gradient descent to local minima. The addition of isotropic noise can significantly slow down the convergence rate. In contrast, our setting is deterministic and corresponds to the simplest possible discrete-time implementation of gradient descent.

Numerous curvature-based optimization techniques have been developed in order to circumvent saddle points (e.g., trust-region methods [5, 28], modified Newton's method with curvilinear line search [17], cubic regularized Newton's method [19], and saddle-free Newton methods [6]). Unlike gradient descent, these methods have superlinear per-iteration implementation costs, making them impractical for high dimensional settings.

Gradient descent with carefully chosen initial conditions can bypass the problem of local minima altogether and converge to the global minimum for many practical non-convex optimization settings (e.g., dictionary learning [1], latent-variable models [29], matrix completion [9], and phase retrieval [2]). In contrast, we focus on the performance of gradient descent under generic initial conditions. Finally, some recent work has been focusing on the connections between stability and efficiency of fixed points in non-convex optimization (e.g., Gaussian random fields [4]).

Gradient-like dynamics, where the dynamic moves towards states of decreased cost but without necessarily moving in the direction of steepest decrease, is a generalization of gradient dynamics that arise in a number of applications including game theory and mathematical biology. Similar arguments about convergence to local minima for almost all initial conditions have been argued [10, 20, 13] for (variants of) replicator dynamics and multiplicative weights update algorithms when applied to games where the incentives of all agents are closely aligned.[5] From the perspective of biology and specifically evolution, (variants of) replicator/MWUA [3, 16] capture standard models of the evolution of the frequencies of different genotypes within a species (preferential survival of the fittest). By analyzing the properties of local minimum energy states we can derive completely different conclusions about the long term system behavior (in terms e.g., of the resulting genetic diversity) from the ones that follow from analyzing all saddle points [13]. In fact, understanding the properties of local minima raises interesting computational complexity questions [15]. Finally, examining the stability properties of equilibria can help us capture quantitatively the long term behavior of biologically inspired gradient-like systems even under time-evolving fitness landscapes [14]. Given the emergent overlapping interests between these areas and (non-convex) optimization theory, it seems that novel opportunities for cross-fertilization between these research communities arise.

1.2 Organization

In Section 2, we introduce the notation and definitions used throughout the paper and state formally our main theorems. In Section 3, we prove our results establishing the negligible probability of converging to saddle points, addressing the possibility of continuums

[5] Such games are known as potential/congestion games [10] and correspond to games where all agents act as if they share a common cost/potential function that they are trying to minimize.

of equilibria, forward-invariant subspaces, and establishing an upper bound on the step-size. In Section 4, we produce several examples showcasing the effectiveness of our methods. Finally, we conclude in Section 5 by suggesting directions for future work.

2 Preliminaries

Notation: We use boldface letters, e.g., $\mathbf{x}$, to denote column vectors. We denote by $\text{sp}(A)$, $\|A\|_2$ the spectral radius and spectral norm of a symmetric matrix A respectively. We also use $\|\mathbf{x}\|_2$ for the ℓ_2 norm of vector $\mathbf{x}$. By $\nabla^2 f(\mathbf{x})$ we denote the Hessian of a twice differentiable function $f : \mathcal{E} \to \mathbb{R}$, for some set $\mathcal{E} \subseteq \mathbb{R}^N$.

Assume a minimization problem of the form $\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})$ where $f : \mathbb{R}^N \to \mathbb{R}$ is a twice continuously differentiable function. Gradient descent is one of the most well-known algorithms (discrete dynamical system) to attack this generic optimization problem. It is defined by the equations below:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k),$$

or equivalently $\mathbf{x}_{k+1} = g(\mathbf{x}_k)$ with $g(\mathbf{x}) = \mathbf{x} - \alpha\nabla f(\mathbf{x})$, $g : \mathbb{R}^N \to \mathbb{R}^N$ and $\alpha > 0$.

It is easy to see that the fixed points of the dynamical system $\mathbf{x}_{k+1} = g(\mathbf{x}_k)$ are exactly the points $\mathbf{x}$ so that $\nabla f(\mathbf{x}) = \mathbf{0}$, called *critical points or equilibria*. The set of local minima of f is a subset of the set of critical points of f. These two sets do not coincide and this poses a serious obstacle for proving strong theoretical guarantees for gradient descent, since the dynamics may converge to a critical point which is not a local minimum, called a *saddle point*.

Lee et al. [12] argue, under technical conditions which include the assumption of isolated critical points, that the set of initial conditions that converge to *strict saddle points* is a zero measure set (for definition of strict saddle, see Definition 1). The paper leaves as an open question whether the condition of isolated equilibria is necessary. We prove that the set of initial conditions that converge to a strict saddle point is a zero measure set even in the case of non-isolated critical points[6]. Furthermore, one of the conditions for f is that ∇f is globally Lipschitz, which implies that the second derivative of f is bounded, i.e., there exists a $\beta > 0$ such that for all $\mathbf{x}$ we have $\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq \beta$. However, even third degree polynomial functions are not globally Lipschitz. We provide a theorem which can circumvent this assumption as long as the domain $\mathcal{S}$ is *forward or positively invariant* with respect to g, i.e., $g(\mathcal{S}) \subseteq \mathcal{S}$. Finally, we provide an easy upper bound on the step-size α, via eigenvalue analysis of the Jacobian of g, i.e., $I - \alpha\nabla^2 f(\mathbf{x})$.

Below we give some necessary definitions as appeared in Lee et al. [12].

▶ **Definition 1.**

- A point $\mathbf{x}^*$ is a critical point of f if $\nabla f(\mathbf{x}^*) = \mathbf{0}$. We denote by $C = \{\mathbf{x} : \nabla f(\mathbf{x}) = \mathbf{0}\}$ the set of critical points (can be uncountably many).
- A critical point $\mathbf{x}^*$ is isolated if there is a neighborhood U around $\mathbf{x}^*$ and $\mathbf{x}^*$ is the only critical point in U.[7] Otherwise is called non-isolated.
- A critical point $\mathbf{x}^*$ of f is a saddle point if for all neighborhoods U around $\mathbf{x}^*$ there are $\mathbf{y}, \mathbf{z} \in U$ such that $f(\mathbf{z}) \leq f(\mathbf{x}^*) \leq f(\mathbf{y})$.

[6] Our arguments hence allow for cost functions f's with uncountably many critical points.

[7] If the critical points are isolated then they are countably many or finite.

- A critical point $\mathbf{x}^*$ of f is a strict saddle if $\lambda_{\min}(\nabla^2 f(\mathbf{x}^*)) < 0$ (minimum eigenvalue of matrix $\nabla^2 f(\mathbf{x}^*)$ is negative).
- A set $\mathcal{S}$ is called forward or positively invariant with respect to some function $h : \mathcal{E} \to \mathbb{R}^N$ with $\mathcal{S} \subseteq \mathcal{E} \subseteq \mathbb{R}^N$ if $h(\mathcal{S}) \subseteq \mathcal{S}$.

2.1 Main Results

In [12], the steps of the proof of their result are the following: Under the regularity assumption that ∇f is globally Lipschitz with some Lipschitz constant L, Lee et al. are able to show that $g(\mathbf{x}) = \mathbf{x} - \alpha \nabla f(\mathbf{x})$ is a diffeomorphism for $\alpha < 1/L$. Afterwards, using the center-stable manifold theorem (see theorem 8), they show that the set of initial conditions so that g converges to saddle points has measure zero under the assumption that the critical points are isolated. We generalize their result for non-isolated critical points, answering one of their open questions (see also the example in Section 4.1, where there is a line of critical points).

▶ **Theorem 2.** [Non-isolated] *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a twice continuously differentiable function and $\sup_{\mathbf{x} \in \mathbb{R}^N} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L < \infty$. The set of initial conditions $\mathbf{x} \in \mathbb{R}^N$ so that gradient descent with step-size $0 < \alpha < 1/L$ converges to a strict saddle point is of (Lebesgue) measure zero, without the assumption that critical points are isolated.*

We can prove a stronger version of the theorem above, circumventing the globally Lipschitz condition for domains which are forward invariant (see also the example in Section 4.2).

▶ **Theorem 3.** [Non-isolated, forward invariant] *Let $f : \mathcal{S} \to \mathbb{R}$ be twice continuously differentiable in an open convex set $\mathcal{S} \subseteq \mathbb{R}^N$ and $\sup_{\mathbf{x} \in \mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L < \infty$. If $g(\mathcal{S}) \subseteq \mathcal{S}$ (where $g(\mathbf{x}) = \mathbf{x} - \alpha \nabla f(\mathbf{x})$) then the set of initial conditions $\mathbf{x} \in \mathcal{S}$ so that gradient descent with step-size $0 < \alpha < 1/L$ converges to a strict saddle point is of (Lebesgue) measure zero, without the assumption that critical points are isolated.*

Finally, via eigenvalue analysis of $I - \alpha \nabla^2 f(\mathbf{x})$, we can find upper bounds on the step-size of gradient descent. A straightforward theorem is the following:

▶ **Theorem 4.** [Upper bound on step-size] *Let f be twice continuously differentiable in an open set $\mathcal{S} \subseteq \mathbb{R}^N$ and $\mathcal{C}^*$ be the set of local minima. Assume also that $\gamma < \inf_{\mathbf{x} \in \mathcal{C}^*} \left\|\nabla^2 f(\mathbf{x})\right\|_2 < \infty$. A necessary condition so that gradient descent converges to local minima for all but (Lebesgue) measure zero initial conditions in $\mathcal{S}$ is that the step-size satisfies $\alpha < \frac{2}{\gamma}$.*

3 Proving the theorems

Before we proceed with the proofs, let us argue that Theorem 3 is a generalization of Theorem 2. This can be checked by setting $\mathcal{S} := \mathbb{R}^N$ and observing that $g(\mathbb{R}^N) \subseteq \mathbb{R}^N$. We continue with the proofs of Theorems 3 and 4.

3.1 Proof of Theorem 3

In this section, we prove Theorem 3. We start by showing (for completeness) that the assumptions of Theorem 3 imply that $\nabla f(\mathbf{x})$ is Lipschitz in $\mathcal{S}$.

▶ **Lemma 5.** *Let $f : \mathcal{S} \to \mathbb{R}$ where $\mathcal{S}$ is an open convex set and f be twice continuously differentiable in $\mathcal{S}$. Also assume that $\sup_{\mathbf{x} \in \mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L < \infty$. Then ∇f satisfies the Lipschitz condition in $\mathcal{S}$ with Lipschitz constant L.*

Proof. Let $\mathbf{x},\mathbf{y} \in \mathcal{S}$ (column vectors) and define the function $H : [0,1] \to \mathbb{R}^N$ as $H(t) = \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$. By the chain rule we get that $H'(t) := \frac{dH}{dt} = (\nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))) \cdot (\mathbf{y}-\mathbf{x})$. It holds that

$$\begin{aligned} \left\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\right\|_2 = \left\|\int_0^1 H'(t)dt\right\|_2 &\leq \int_0^1 \left\|H'(t)\right\|_2 dt \\ &= \int_0^1 \left\|(\nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})))(\mathbf{y}-\mathbf{x})\right\|_2 dt \\ &\leq \int_0^1 \left\|\nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))\right\|_2 \left\|\mathbf{y}-\mathbf{x}\right\|_2 dt \\ &\leq \int_0^1 L \left\|\mathbf{y}-\mathbf{x}\right\|_2 dt = L \left\|\mathbf{y}-\mathbf{x}\right\|_2 . \end{aligned}$$

◀

▶ Remark. From Schwarz's theorem we get that $\nabla^2 f(\mathbf{x})$ is symmetric for $\mathbf{x} \in \mathcal{S}$, hence $\left\|\nabla^2 f(\mathbf{x})\right\|_2 = \mathrm{sp}(\nabla^2 f(\mathbf{x}))$.

The assumption that $\sup_{\mathbf{x}\in\mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L < \infty$ implies that $\nabla f(x)$ is Lipschitz with constant L in the convex set $\mathcal{S}$, as stated by Lemma 5. We show that the converse holds as well, i.e., the Lipschitz condition for $\nabla f(\mathbf{x})$ with constant L in the main theorem in Lee et al. implies $\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$ for all $\mathbf{x} \in \mathcal{S}$ and hence the assumption in our Theorems 2, 3 that $\sup_{\mathbf{x}\in\mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$ is satisfied.

▶ **Lemma 6.** *Let $f : \mathcal{S} \to \mathbb{R}$ where $\mathcal{S} \subseteq \mathbb{R}^N$ is an open convex set and f is twice continuously differentiable in $\mathcal{S}$. Assume $\nabla f(\mathbf{x})$ is Lipschitz with constant L in $\mathcal{S}$ then it holds $\sup_{\mathbf{x}\in\mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$.*

Proof. Fix an $\epsilon > 0$. By Taylor's theorem since f is twice differentiable with respect to some point $\mathbf{x}$ it holds that

$$\begin{aligned} \left\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\right\|_2 &\geq \left\|(\nabla^2 f(\mathbf{x}))(\mathbf{y}-\mathbf{x})\right\|_2 - o(\left\|\mathbf{y}-\mathbf{x}\right\|_2) \\ &\geq \left\|(\nabla^2 f(\mathbf{x}))(\mathbf{y}-\mathbf{x})\right\|_2 - \epsilon \left\|\mathbf{y}-\mathbf{x}\right\|_2 \end{aligned}$$

for $\mathbf{y}$ sufficiently close to $\mathbf{x}$ (depends on ϵ). Therefore under the Lipschitz assumption we get that there exists a closed neighborhood $U(\epsilon)$ of $\mathbf{x}$ so that for all $\mathbf{y} \in U$ we get

$$\left\|(\nabla^2 f(\mathbf{x}))(\mathbf{x}-\mathbf{y})\right\|_2 \leq \left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|_2 + \epsilon \left\|\mathbf{y}-\mathbf{x}\right\|_2 \leq (L+\epsilon) \left\|\mathbf{x}-\mathbf{y}\right\|_2 . \tag{1}$$

We consider a closed ball B subset of U, with center $\mathbf{x}$ and radius r (in ℓ_2) and set $\mathbf{z} = \mathbf{x}-\mathbf{y}$. It is true that $\left\|\nabla^2 f(\mathbf{x})\right\|_2 = \sup_{\left\|\mathbf{z}\right\|_2 = r} \frac{\left\|(\nabla^2 f(\mathbf{x}))\mathbf{z}\right\|_2}{\left\|\mathbf{z}\right\|_2}$ by definition of spectral norm, scaled so that the length of the vectors is exactly r. Using 1 we get that $\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L+\epsilon$. Since ϵ is arbitrary, we get that $\left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$. We conclude that $\sup_{\mathbf{x}\in\mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$. ◀

Lemmas 5 and 6 (which are provided for completeness) show that the smoothness assumptions in Lee et al. paper are equivalent to ours. We use the condition on the spectral norm of the matrix $\nabla^2 f(\mathbf{x})$ so that we can work with the eigenvalues in our theorems (e.g., in Remark 3.1 the spectral norm coincides with spectral radius for $\nabla^2 f(\mathbf{x})$). Below we prove that the update rule of gradient descent, i.e., function g is a diffeomorphism under the assumptions of Theorem 3 (similar approach appeared in [12]).

▶ **Lemma 7.** *Under the assumptions of Theorem 3, function g is a diffeomorphism in $\mathcal{S}$.*

Proof. First we prove that g is a injective. We follow the same argument as in [12]. Suppose $g(\mathbf{y}) = g(\mathbf{x})$, thus $\mathbf{y} - \mathbf{x} = \alpha(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))$. We assume that $\mathbf{x} \neq \mathbf{y}$ and we will reach contradiction. From Lemma 5 we get $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L \|\mathbf{y} - \mathbf{x}\|_2$ and hence $\|\mathbf{x} - \mathbf{y}\|_2 \leq \alpha L \|\mathbf{y} - \mathbf{x}\|_2 < \|\mathbf{y} - \mathbf{x}\|_2$ since $\alpha L < 1$ (contradiction).

We continue by showing that g is a local diffeomorphism. Observe that the Jacobian of g is $I - \alpha\nabla^2 f(\mathbf{x})$. It suffices to show that $\alpha\nabla^2 f(\mathbf{x})$ has no eigenvalue which is 1, because this implies matrix $I - \alpha\nabla^2 f(\mathbf{x})$ is invertible. As long as $I - \alpha\nabla^2 f(\mathbf{x})$ is invertible, from Inverse Function Theorem (see [27]) follows that g is a local diffeomorphism. Finally, since g is injective, the inverse g^{-1} is well defined and since g is a local diffeomorphism in $\mathcal{S}$, it follows that g^{-1} is smooth in $\mathcal{S}$. Therefore g is a diffeomorphism.

Let λ be an eigenvalue of $\nabla^2 f(\mathbf{x})$. Then $|\lambda| \leq \text{sp}(\nabla^2 f(\mathbf{x})) = \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$ where the equality comes from Remark 3.1 and first and last inequalities are satisfied by assumption. Therefore $\alpha\nabla^2 f(\mathbf{x})$ has as eigenvalue $\alpha\lambda$ and $|\alpha\lambda| \leq \alpha L < 1$. Thus all eigenvalues of $\alpha\nabla^2 f(\mathbf{x})$ are less than 1 in absolute value and the proof is complete. ◀

Proof of Theorem 3. We will use the Center-stable manifold theorem since $g(\mathbf{x}) = \mathbf{x} - \alpha\nabla f(\mathbf{x})$ is a diffeomorphism, where $\sup_{\mathbf{x}\in\mathcal{S}} \left\|\nabla^2 f(\mathbf{x})\right\|_2 \leq L$ and $\alpha < 1/L$. A modification of the proof of Theorem 3 appeared in [20] and [13] for replicator dynamics (not gradient descent).

▶ **Theorem 8** (Center and Stable Manifolds, p. 65 of [26]). *Let $\mathbf{p}$ be a fixed point for the C^r local diffeomorphism $h : U \to \mathbb{R}^n$ where $U \subset \mathbb{R}^n$ is an open neighborhood of $\mathbf{p}$ in $\mathbb{R}^n$ and $r \geq 1$. Let $E^s \oplus E^c \oplus E^u$ be the invariant splitting of $\mathbb{R}^n$ into generalized eigenspaces of $Dh(\mathbf{p})$[8] corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. To the $Dh(\mathbf{p})$ invariant subspace $E^s \oplus E^c$ there is an associated local h invariant C^r embedded disc W^{sc}_{loc} of dimension dim$(E^s \oplus E^c)$, and ball B around $\mathbf{p}$ such that:*

$$h(W^{sc}_{loc}) \cap B \subset W^{sc}_{loc}.\ \textit{If } h^n(\mathbf{x}) \in B \textit{ for all } n \geq 0,\ \textit{ then } \mathbf{x} \in W^{sc}_{loc}. \tag{2}$$

From this point on our approach deviates significantly from that of [12] and new ideas and tools need to be introduced.

Let $\mathbf{r}$ be a critical point of function $f(\mathbf{x})$ and $B_{\mathbf{r}}$ be the (open) ball that is derived from Theorem 8. We consider the union of these balls

$$A = \cup_{\mathbf{r}} B_{\mathbf{r}}.$$

The following property for $\mathbb{R}^N$ holds:

▶ **Theorem 9** (Lindelőf's lemma [8]). *For every open cover there is a countable subcover.*

Therefore due to Lindelőf's lemma, we can find a countable subcover for A, i.e., there exist fixed points $\mathbf{r}_1, \mathbf{r}_2, \ldots$ such that $A = \cup_{m=1}^{\infty} B_{\mathbf{r}_m}$. If gradient descent converges to a strict saddle point, starting from a point $\mathbf{v} \in \mathcal{S}$, there must exist a t_0 and m so that $g^t(\mathbf{v}) \in B_{\mathbf{r}_m}$ for all $t \geq t_0$. From Theorem 8 we get that $g^t(\mathbf{v}) \in W^{sc}_{loc}(\mathbf{r}_m) \cap \mathcal{S}$ where we used the fact that $g(\mathcal{S}) \subseteq \mathcal{S}$ (from assumption forward invariant), namely the trajectory remains in $\mathcal{S}$ for all times [9]. By setting $D_1(\mathbf{r}_m) = g^{-1}(W^{sc}_{loc}(\mathbf{r}_m) \cap \mathcal{S})$ and $D_{i+1}(\mathbf{r}_m) = g^{-1}(D_i(\mathbf{r}_m) \cap \mathcal{S})$ we get

[8] Jacobian of h evaluated at $\mathbf{p}$.

[9] $W^{sc}_{loc}(\mathbf{r}_m)$ denotes the center stable manifold of fixed point $\mathbf{r}_m$

that $\mathbf{v} \in D_t(\mathbf{r}_m)$ for all $t \geq t_0$. Hence the set of initial points in $\mathcal{S}$ so that gradient descent converges to a strict saddle point is a subset of

$$P = \cup_{m=1}^{\infty} \cup_{t=0}^{\infty} D_t(\mathbf{r}_m)). \tag{3}$$

Since $\mathbf{r}_m$ is strict saddle point, the Jacobian $I - \alpha\nabla^2 f(\mathbf{x})$ has an eigenvalue greater than 1, namely the dimension of the unstable eigenspace satisfies $dim(E^u) \geq 1$, and therefore the dimension of $W^{sc}_{loc}(\mathbf{r}_m)$ is at most $N - 1$. Thus, the set $W^{sc}_{loc}(\mathbf{r}_m) \cap \mathcal{S}$ has Lebesgue measure zero in $\mathbb{R}^N$. Finally since g is a diffeomorphism (from Lemma 7), g^{-1} is continuously differentiable and thus it is locally Lipschitz (see [22] p.71). Therefore using Lemma 10 which we state below, g^{-1} preserves the null-sets and hence (by induction) $D_i(\mathbf{r}_m)$ has measure zero for all i. Thereby we get that P is a countable union of measure zero sets, i.e., is measure zero as well and the claim of Theorem 3 follows. ◀

▶ **Lemma 10.** *Let $h : \mathcal{S} \to \mathbb{R}^m$ be a locally Lipschitz function with $\mathcal{S} \subseteq \mathbb{R}^m$ then h is null-set preserving, i.e., for $E \subset \mathcal{S}$ if E has measure zero then $h(E)$ has also measure zero.*

Proof. The lemma is quite standard, but we provide a proof for completeness. Let B_γ be an open ball such that $\|h(\mathbf{y}) - h(\mathbf{x})\| \leq K_\gamma \|\mathbf{y} - \mathbf{x}\|$ for all $\mathbf{x}, \mathbf{y} \in B_\gamma$. We consider the union $\cup_\gamma B_\gamma$ which cover $\mathbb{R}^m$ by the assumption that h is locally Lipschitz. By Lindelőf's lemma we have a countable subcover, i.e., $\cup_{i=1}^{\infty} B_i$. Let $E_i = E \cap B_i$. We will prove that $h(E_i)$ has measure zero. Fix an $\epsilon > 0$. Since $E_i \subset E$, we have that E_i has measure zero, hence we can find a countable cover of open balls $C_1, C_2, ...$ for E_i, namely $E_i \subset \cup_{j=1}^{\infty} C_j$ so that $C_j \subset B_i$ for all j and also $\sum_{j=1}^{\infty} \mu(C_j) < \frac{\epsilon}{K_i^m}$. Since $E_i \subset \cup_{j=1}^{\infty} C_j$ we get that $h(E_i) \subset \cup_{j=1}^{\infty} h(C_j)$, namely $h(C_1), h(C_2), ...$ cover $h(E_i)$ and also $h(C_j) \subset h(B_i)$ for all j. Assuming that ball $C_j \equiv B(\mathbf{x}, r)$ (center $\mathbf{x}$ and radius r) then it is clear that $h(C_j) \subset B(h(\mathbf{x}), K_i r)$ (h maps the center $\mathbf{x}$ to $h(\mathbf{x})$ and the radius r to $K_i r$ because of Lipschitz assumption). But $\mu(B(h(\mathbf{x}), K_i r)) = K_i^m \mu(B(\mathbf{x}, r)) = K_i^m \mu(C_j)$, therefore $\mu(h(C_j)) \leq K_i^m \mu(C_j)$ and so we conclude that

$$\mu(h(E_i)) \leq \sum_{j=1}^{\infty} \mu(h(C_j)) \leq K_i^m \sum_{j=1}^{\infty} \mu(C_j) < \epsilon$$

Since ϵ was arbitrary, it follows that $\mu(h(E_i)) = 0$. To finish the proof, observe that $h(E) = \cup_{i=1}^{\infty} h(E_i)$ therefore $\mu(h(E)) \leq \sum_{i=1}^{\infty} \mu(h(E_i)) = 0$. ◀

A straightforward application of Theorem 3 is the following:

▶ **Corollary 11.** *Assume that the conditions of Theorem 3 are satisfied and all saddle points of f are strict. Additionally, let ν be a prior measure with support $\mathcal{S}$ which is absolutely continuous with respect to Lebesgue measure, and assume $\lim_{k\to\infty} g^k(x)$ exists[10] for all $\mathbf{x}$ in $\mathcal{S}$. Then*

$$\mathbb{P}_\nu[\lim_k g^k(\mathbf{x}) = \mathbf{x}^*] = 1,$$

where $\mathbf{x}^*$ is a local minimum.

Proof. Since the set of initial conditions whose limit point is a (strict) saddle point is a measure zero set and we have assumed $\lim_{k\to\infty} g^k(x)$ exists for all initial conditions in $\mathcal{S}$ then the probability of converging to a local minimizer is 1. ◀

[10] g^k denotes the composition of g with itself k times.

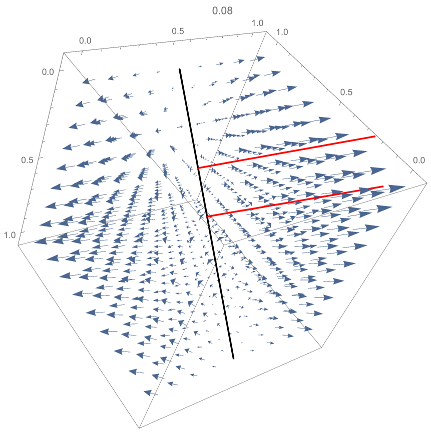

Figure 1 Example that satisfies the assumptions of Theorem 2. The black line represent critical points of f, all of which are strict. The red lines correspond to diverging trajectories of gradient descent with small step size.

▶ Remark. Arguing that $\lim_k g^k(\mathbf{x})$ exists follows from standard arguments in several settings of interest (e.g for analytic functions f that satisfy (Lojasiewicz Gradient Inequality)), see paper [12] and references therein.

The importance of Theorem 3 will become clear in the examples of Section 4. Specifically, in the example of Section 4.2, the function is not globally Lipschitz (we use the example that appears in [12]), nevertheless Theorem 3 applies and thus we have convergence to local minimizers with probability 1. In the example of Section 4.1 we see that simple functions may have non-isolated critical points.

3.2 Proof of Theorem 4

Proof. We proceed by contradiction. Consider any local minimum $\mathbf{x}^*$, and by assumption we get that $\text{sp}(\nabla^2 f(\mathbf{x}^*)) > \gamma$. Let $\alpha \geq \frac{2}{\gamma}$. Therefore the Jacobian $I - \alpha\nabla^2 f(\mathbf{x}^*)$ of g at $\mathbf{x}^*$ has spectral radius greater than 1 since $\text{sp}(I - \alpha\nabla^2 f(\mathbf{x}^*)) \geq \text{sp}(\alpha\nabla^2 f(\mathbf{x}^*)) - 1 > \alpha\gamma - 1 \geq 1$. This implies that the fixed point $\mathbf{x}^*$ of g is (Lyapunov) unstable. Since this is true for every local minimum, it cannot be true that gradient descent converges with probability 1 to local minima. ◀

4 Examples

4.1 Example for non-isolated critical points

Consider the simple example of the cost function $f : \mathbb{R}^3 \to \mathbb{R}$ with $f(x, y, z) = 2xy + 2xz - 2x - y - z$. Its gradient is $\nabla(f) = (2y + 2z - 2, 2x - 1, 2x - 1)$. Naturally, its saddle points correspond exactly to the line $(1/2, w, 1 - w)$ for $w \in \mathbb{R}$ and by computing their (common) eigenvalues we establish that they are all strict saddles (their minimum eigenvalue is $-2\sqrt{2}$). As we expect from our analysis effectively no trajectories converge to them (instead the value of practically all trajectories goes to $-\infty$). We plot in red some sample trajectories for small enough step sizes, starting in the local neighborhood of the equilibrium set.

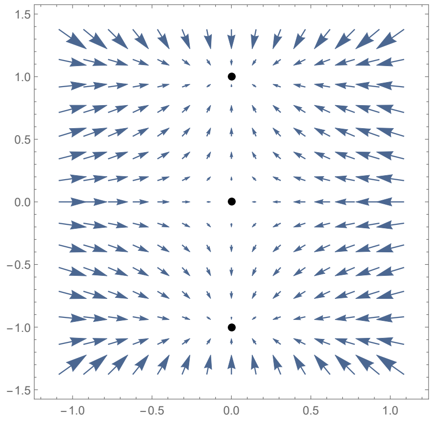

Figure 2 Example that satisfies the assumptions of Theorem 3. The three black dots represent the critical points. Function f is not Lipschitz.

4.2 Example for forward invariant set

We use the same function as in Lee et al. $f(x,y) = \frac{x^2}{2} + \frac{y^4}{4} - \frac{y^2}{2}$. As argued in previous sections, f is not globally Lipschitz so the main result in [12] cannot be applied here. We will use our Theorem 3 which talks about forward invariant domains.

The critical points of f are $(0,0), (0,1), (0,-1)$. $(0,0)$ is a strict saddle point and the other two are local minima. Observe that the Hessian $\nabla^2 f(x,y)$ is

$$J = \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix}.$$

For $\mathcal{S} = (-1,1) \times (-2,2)$, so we get that $\sup_{(x,y)\in\mathcal{S}} \left\|\nabla^2 f(x,y)\right\|_2 \leq 11$ (for $y = 2$ gets the maximum value). We choose $\alpha = \frac{1}{12} < \frac{1}{11}$, and we have $g(x,y) = ((1-\alpha)x, (1+\alpha)y - \alpha y^3) = (\frac{11x}{12}, \frac{13y}{12} - \frac{y^3}{12})$. It is not difficult to see that $g(\mathcal{S}) \subseteq \mathcal{S}$ (easy calculations). The assumptions of Theorem 3 are satisfied, hence it is true that the set of initial conditions in $\mathcal{S}$ so that gradient descent converges to $(0,0)$ has measure zero. Moreover, by Corollary 11 it holds that if the initial condition is taken (say) uniformly at random in $\mathcal{S}$, then gradient descent converges to $(0,1), (0,-1)$ with probability 1. The figure below makes the claim clear, i.e. the set of initial conditions so that gradient descent converges to $(0,0)$ lie on the axis $y = 0$, which is of measure zero in $\mathbb{R}^2$. For all other starting points, gradient descent converges to local minima. Finally, from the figure one can see that $\mathcal{S}$ is forward invariant.

4.3 Example for step-size

We use the same function as in the previous example. Observe that for $(0,0), (0,1), (0,-1)$ we have that the spectral radius of $\nabla^2 f$ is $1, 2, 2$ respectively (so the minimum of all is 1). We choose $\alpha \geq 2$ and we get that $g(x,y) = (-x, 3y - 2y^3)$. It is not hard to see that gradient descent does not converge (in the first coordinate function g cycles between x and $-x$).

5 Conclusion

Our positive result cannot be improved without making explicit assumptions on the structure of the cost function f nor using beneficial random noise/well chosen initial conditions (as far as convergence is concerned, speed of convergence is not in the scope of this paper). Naturally, all these directions are of key interest and are the object of recent work (see section 1.1). Keeping up with this simplest, deterministic implementation of gradient descent a natural hypothesis is that (in settings of practical interest) it converges not only to local minimizers but moreover the size of the region of attraction of each local minimizer is in a sense directly proportional to its quality.

Recently, in [20] there has been some progress in proving such statements in non-convex gradient-like systems that arise from learning in games. In such settings, (stable) fixed points correspond to Nash equilibria, but instead of having the typical system performance being dominated by the worst case Nash equilibria (as Price of Anarchy suggests) the regions of attractions of such bad (social) states prove to be minimal and the system works near optimally on average (given uniformly random initial conditions). Extending such statements to actual gradient-dynamics as well as comparing the average case performance of different heuristics even in restricted settings is a fascinating question that could shed more light into the in-many-cases surprising efficiency of the gradient descent method.

Acknowledgements. We are grateful to Prasad Tetali, Jason D. Lee, Max Simchowitz, Michael I. Jordan and Benjamin Recht for their support and helpful discussions and suggestions as well as for the elucidating blog article at `http://www.offconvex.org` on their elegant work, which inspired our investigation. We are also thankful to Nisheeth Vishnoi, on whose blog the article appeared, for pointing it out to us. This work was completed while Ioannis Panageas was a PhD student at Georgia Institute of Technology. Ioannis Panageas would like to acknowledge NSF EAGER award grants CCF-1415496, CCF-1415498 and a MIT-SUTD postdoctoral fellowship. Georgios Piliouras would like to acknowledge SUTD grant SRG ESD 2015 097 and MOE AcRF Tier 2 Grant 2016-T2-1-170. Part of the work was completed while Ioannis Panageas and Georgios Piliouras were visiting scientists at the Simons Institute for the Theory of Computing.

References

1 Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *28th Conference on Learning Theory (COLT)*, pages 113–149, 2015.

2 Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

3 Erick Chastain, Adi Livnat, Christos Papadimitriou, and Umesh Vazirani. Algorithms, games, and evolution. *Proceedings of the National Academy of Sciences (PNAS)*, 111(29):10620–10623, 2014.

4 Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. *arXiv preprint arXiv:1412.0233*, 2014.

5 Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*, volume 1. Siam, 2000.

6 Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems (NIPS)*, pages 2933–2941, 2014.

7 Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. *arXiv preprint arXiv:1503.02101*, 2015.

8 John L. Kelley. *General Topology.* Springer, 1955.

9 Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. *IEEE International Symposium on Information Theory (ISIT)*, pages 324–328, 2009.

10 Robert Kleinberg, Georgios Piliouras, and Eva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. *Symposium on Theory of Computing (STOC)*, pages 533–542, 2009.

11 Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems (NIPS)*, pages 556–562, 2001.

12 Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. *Conference on Learning Theory (COLT)*, 2016.

13 Ruta Mehta, Ioannis Panageas, and Georgios Piliouras. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. *Innovations in Theoretical Computer Science (ITCS)*, 2015.

14 Ruta Mehta, Ioannis Panageas, Georgios Piliouras, Prasad Tetali, and Vijay V. Vazirani. Mutation, Sexual Reproduction and Survival in Dynamic Environments. *Innovations in Theoretical Computer Science (ITCS)*, 2017.

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

16 Reshef Meir and David Parkes. On sex, evolution, and the multiplicative weights update algorithm. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 929–937. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

17 Jorge J Moré and Danny C Sorensen. On the use of directions of negative curvature in a modified newton method. *Mathematical Programming*, 16(1):1–20, 1979.

18 Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science and Business Media, 2004.

19 Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

20 Ioannis Panageas and Georgios Piliouras. Average case performance of replicator dynamics in potential games via computing regions of attraction. *17th ACM Conference on Economics and Computation (EC)*, 2016.

21 Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, pages 698–712, 1990.

22 Lawrence Perko. *Differential Equations and Dynamical Systems.* Springer, 3nd. edition, 1991.

23 A Ravindran, Gintaras Victor Reklaitis, and Kenneth Martin Ragsdell. *Engineering optimization: methods and applications.* John Wiley & Sons, 2006.

24 Levent Sagun, Leon Bottou, and Yann LeCun. Singularity of the hessian in deep learning. *arXiv preprint arXiv:1611.07476*, 2016.

25 William H Sandholm. Evolutionary game theory. In *Encyclopedia of Complexity and Systems Science*, pages 3176–3205. Springer, 2009.

26 Michael Shub. *Global Stability of Dynamical Systems.* Springer-Verlag, 1987.

27 Michael Spivak. *Calculus On Manifolds: A Modern Approach To Classical Theorems Of Advanced Calculus.* Addison-Wesley, 1965.

28 Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *arXiv preprint arXiv:1511.04777*, 2015.

29 Yuchen Zhang, Xi Chen, Denny Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Advances in Neural Information Processing Systems (NIPS)*, pages 1260–1268, 2014.