Revisiting Normalized Gradient Descent: Fast Evasion of Saddle Points

Ryan Murray∗‡, Brian Swenson ∗†, and Soummya Kar†

Abstract—The paper considers normalized gradient descent (NGD), a natural modification of classical gradient descent (GD) in optimization problems. It is shown that, contrary to GD, NGD escapes saddle points “quickly.” A serious shortcoming of GD is that it can take arbitrarily long to escape from the neighborhood of a saddle point. In practice, this issue can significantly slow the convergence of GD, particularly in high dimensional nonconvex problems. The paper focuses on continuous-time optimization dynamics. It is shown that (i) NGD “almost never” converges to saddle points and (ii) the time required for NGD to escape from a ball of radius r about a saddle point x∗ is at most \(\frac{\kappa}{\sqrt{r}}\), where \(\kappa\) is the condition number of the Hessian of \(f\) at \(x^\ast\). As a simple application of these results, a global convergence-time bound is established for NGD under mild assumptions.

I. INTRODUCTION

Given a differentiable function \(f : \mathbb{R}^d \rightarrow \mathbb{R}\), the canonical first-order optimization procedure is the method of gradient descent (GD). In continuous time, GD is defined by the differential equation

\[
\dot{x} = -\nabla f(x).
\]  

(1)

Discrete-time GD, commonly used in applications, is merely a sample and hold (or Euler) discretization (with diminishing step size) of the differential equation (1), and the properties of discrete- and continuous-time GD are closely related [1]–[3]. Owing to their simplicity and ease of implementation, GD and related first-order optimization procedures are popular in practice, particularly in large-scale problems where second-order information such as the Hessian can be costly to compute [4]. When the objective function \(f\) is convex, GD can be both practical and effective as an optimization procedure. However, when \(f\) is non-convex, GD can perform poorly in practice, even when the goal is merely to find a local minimum.

The underlying issue is the presence of saddle points in non-convex functions; the gradient \(\nabla f(x)\) vanishes near saddle points, which causes GD to “stall” in neighboring regions [5] (see also Section III-A). This both slows the overall convergence rate and makes detection of local minima difficult. The detrimental effects of this issue become particularly severe in high-dimensional problems where the number of saddle points may proliferate. For example, recent work [5] showed that in some high-dimensional problems of interest, the number of saddle points increases exponentially relative to the number of local minima, which can dramatically increase the time required for GD to find even a local minimum.

Since first-order dynamics such as GD tend to be relatively simple to implement in large-scale applications, there has been a growing interest in understanding the issue of saddle-point slowdown in non-convex problems and how to overcome it [5]–[10]. This has been a topic of particular interest within the machine learning community where large-scale non-convex optimization and first-order methods are important in many applications [5], [6], [11]–[13].

One intuitively simple method that has been proposed to mitigate this issue is to consider normalized gradient descent (NGD). In continuous time, NGD (originally introduced in [14]) is defined by the differential equation

\[
\dot{x} = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.
\]  

(2)

As with GD, discrete-time NGD [15], [16] is merely a sample and hold discretization of its continuous-time counterpart (2).

The normalized gradient \(\frac{\nabla f(x)}{\|\nabla f(x)\|}\) preserves the direction of the gradient but ignores magnitude. Because \(\nabla f(x)\) does not vanish near saddle points, the intuitive expectation (corroborated by evidence [16]) is that NGD should not slow down in the neighborhood of saddle points and should therefore escape “quickly.”

In this note, our goal is to elucidate the key differences between GD and NGD and, more importantly, give rigorous theoretical justification to the intuition that NGD “escapes saddle points quickly.” We will focus, in this work, on continuous-time descent. From the control perspective, this may be seen as extending the seminal work of [14] by characterizing saddle-point behavior of NGD. From the optimization perspective, focusing on continuous-time dynamics allows us to more easily characterize the fundamental properties of NGD using a wealth of available analysis tools and follows in the spirit of recent works studying optimization processes through the lens of differential equations [17], [18].

We have three main results, which we state informally here:

**Main Result 1** (Theorem 8): Our first main result is to show that NGD (2) can only converge to saddle points from a set of initial conditions with measure zero. We note that this result implies that, generically, NGD only converges to minima of \(f\). However, it provides no guarantees about convergence time or saddle-point escape time. (Indeed, this same result is known to hold for GD, which performs poorly in practice due to saddle-point slow down.)

**Main Result 2** (Theorem 9): Our second main result is to show that NGD always escapes from saddle points “quickly.”

---

*These authors contributed equally.
†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA (brianswe@ece.cmu.edu, soummya@andrew.cmu.edu).
‡Department of Mathematics, Pennsylvania State University, State College, PA, USA (rwm22@psu.edu).

1When we say that a property holds generically for an ordinary differential equation (ODE), we mean that it holds from all initial conditions except, possibly, some set with Lebesgue measure zero.
More precisely, we show that the maximum amount of time a trajectory of NGD can spend in a ball of radius \( r > 0 \) (sufficiently small) about a (non-degenerate) saddle point \( x^* \) is \( 5\sqrt{\kappa r} \), where \( \kappa \) is the condition number of the Hessian of \( f \) at \( x^* \).

We note that this result is independent of the dimension of the problem. In contrast to this, the saddle-point escape time of GD (i.e., the maximum amount of time a trajectory of GD may take to leave a ball of radius \( r \) about a saddle point) is always infinite, independent of the function \( f \), the particular saddle point \( x^* \), or the dimension of the problem. (See Theorem 9 for a precise definition of saddle-point escape time and Remark 5 for a discussion of GD saddle-point escape time.) This is precisely the issue which causes GD to perform poorly in high-dimensional problems with many saddle points.

We remark that studying saddle-point escape time of NGD is challenging due to the discontinuity in the right-hand side of (2) at saddle points. In particular, the system is not amenable to classical analytical techniques. We prove Theorem 9 by studying the rate of “potential energy dissipation” (to use an analogy from physics) of NGD near saddle points. The methods used are flexible and can be applied to a variety of discontinuous dynamical systems (see Remark 13 and Proposition 19).

Main Result 3 (Corollary 18): As our final main result, using the local saddle-point analysis noted above (Theorem 9) we provide a simple global bound on the convergence time of NGD under mild assumptions on \( f \).

Literature Review: Continuous-time NGD dynamics were first introduced by Cortes [14] in the context of distributed multi-agent coordination. In [14] it was shown that NGD converges to critical points of \( f \) in finite time and this result was used to develop distributed gradient coordination algorithms that achieve a desired task in finite time. Our results differ from [14] primarly in that we characterize the saddle-point behavior of NGD, including demonstrating non-convergence to saddle points and providing a strong characterization of saddle-point escape time. Furthermore, our results differ from [14] in that (i) our results show that NGD almost always converges to local minima, rather than just the set of critical points of \( f \), and (ii) [14] considered only local bounds on the convergence time of NGD to local minima. Because we characterize the saddle point behavior of NGD, our results enable global bounds on the convergence time of NGD to minima of non-convex functions (see Corollary 18).

Discrete-time NGD was first introduced by Nesterov [15] and variants have received increasing attention in the optimization and machine learning communities [16], [19]–[21]. The problem of coping with saddle points in non-convex optimization has received significant recent attention—see [5]–[10] and references therein. Of particular relevance to the present work are results dealing with first-order methods. Recent work along these lines includes the following: The work [7] showed that the classical stable manifold theorem implies that gradient descent only converges to minima. The work [9] showed that, even with random initialization, discrete-time GD can take exponential time to escape saddle points. The work [6] showed that noisy discrete-time GD converges to a local minimum in a polynomial number of iterations. Our work differs from [6] (beyond the fact that we study continuous-time dynamics) primarily in that we investigate the role of normalization of the dynamics (rather than noise injection) as a means of accelerating escape from saddle points.

The use of normalization in GD has also been studied in [16], where it was shown that discrete-time NGD with noise injection can outperform GD with noise injection [6] in terms of dimensional dependence and the number of iterations required to reach the basin of a local minimum. Numerical simulations of discrete-time noisy NGD and comparisons with discrete-time noisy GD in several problems of interest were also performed in [16]. Our work differs from [16] in that we study the continuous-time deterministic NGD dynamics (2) (which may be viewed as the mean dynamics of the noise-injected discrete-time NGD [16] as the step size is brought to zero), we characterize the stable manifold for these dynamics near saddle points, and we explicitly characterize the saddle-point escape time.

The work [22] improved on the dimensional dependence of the results of [6] and [16], showing that GD with noise injection can reach the basin of a local minimum in a number of iterations with only polylog dependence on dimension. Our work differs from [22] in that we again study the underlying continuous dynamics and perform an explicit local analysis of the dynamics near saddle points. We demonstrate that the local saddle-point escape time of NGD can be bounded independent of dimension (Theorem 9). Moreover, because we show that NGD is a path-length reparametrization of GD, our results also have implications for classical GD. In particular, Theorem 9 together with Proposition 7 shows that a classical GD trajectory can have at most length \( 5\sqrt{\kappa r} \) (where \( \kappa \) is the condition number of the Hessian of \( f \) at \( x^* \)) before it must exit a ball of radius \( r \) about a saddle point.

Organization: Section II sets up notation. Section III presents a simple example illustrating the salient features of GD and NGD near saddle points. Section IV studies the structural relationship between GD and NGD and presents Theorem 8, which shows generic non-convergence to saddle points. Section V presents Theorem 9, which gives the saddle-point escape-time bound for NGD. Section VI presents a simple global convergence-time bound for NGD (Corollary 18). The proofs of all results are deferred to Section VII. Finally, Section VIII concludes the paper.

II. Preliminaries

Suppose \( f : \mathbb{R}^d \to \mathbb{R} \) is a twice differentiable function. We use the following notation.

- \( \nabla f(x) \) denotes the gradient of \( f \) at \( x \)
- \( D^2 f(x) \) denotes the Hessian of \( f \) at \( x \)
Given a set \( S \subset \mathbb{R}^d \), the closure of \( S \) is given by \( \operatorname{cl}(S) \) and the boundary of \( S \) is given by \( \partial S \).

- \( L^d, d \geq 1 \) denotes the \( d \)-dimensional Lebesgue measure.
- \( B_r(x) \) denotes the open ball of radius \( r \) about \( x \in \mathbb{R}^d \).
- \( \| \cdot \| \) denotes the Euclidean norm.
- \( d(\cdot, \cdot) \) denotes Euclidean distance.
- \( x(t) = \frac{d}{dt} x(t) \).
- For \( A \subset \mathbb{R}^{n \times n} \), \( \sigma(A) \) denotes the spectrum of \( A \).
- \( |\lambda|_{\min}(A) := \min\{ |\lambda| : \lambda \in \sigma(A) \} \)
- \( |\lambda|_{\max}(A) := \max\{ |\lambda| : \lambda \in \sigma(A) \} \).
- The condition number of \( A \) is given by \( \frac{|\lambda|_{\max}(A)}{|\lambda|_{\min}(A)} \).
- \( \text{diag}(\lambda_1, \ldots, \lambda_d) \) gives a \( d \times d \) matrix with \( \lambda_1, \ldots, \lambda_d \) on the diagonal.
- Given \( C > 0 \), \( |D^3 f(x)| < C \) means that \( \left| \frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k} \right| < C \), \( i, j, k = 1, \ldots, d \).

We say that a saddle point \( x^* \) of \( f \) is non-degenerate if \( D^2 f(x^*) \) is non-singular.

For \( k \geq 1 \), let \( C^k \) denote the set of all functions from \( \mathbb{R}^d \) to \( \mathbb{R} \) that are \( k \)-times continuously differentiable. Unless otherwise specified, we will assume the following throughout the paper.

**Assumption 1. The objective function \( f \) is of class \( C^2 \).**

We say that a continuous mapping \( x : I \to \mathbb{R}^d \), over some interval \( I = [0, T) \), \( 0 < T \leq \infty \), is a solution to an ODE with initial condition \( x_0 \) if \( x \in C^1 \), \( x \) satisfies the ODE for all \( t \in I \), and \( x(0) = x_0 \).

Under Assumption 1, there exists a unique solution to (1) on the interval \( I = [0, \infty) \) for every initial condition. On the other hand, a solution \( x \) to (2) with initial condition \( x_0 \) satisfying \( \nabla f(x_0) \neq 0 \), will have a unique solution on some maximal interval of existence \( [0, T) \), where \( T \) is dependent on \( x_0 \) (see [23] for a formal definition of the maximal interval of existence). Practically, for solutions of (2), the maximal interval of existence is the maximal time interval for which a solution \( x \) does not intersect with a critical point of \( f \). When we refer to a solution of (2), we mean the solution defined over its maximal interval of existence.

The following two standard definitions from classical ODE theory will be useful for relating solutions of SGD and GD.

**Definition 2 (Orbit of an ODE).** Let \( x \) be the solution of some ODE on the interval \( [0, T) \) with initial condition \( x(0) = x_0 \). Assume that \( [0, T) \) is the maximal interval on which \( x \) is the unique solution of the ODE with initial value \( x_0 \) (here \( T = \infty \) is permitted). Then the orbit (or forward orbit) corresponding to the initial condition \( x_0 \) is defined to be the set \( \gamma^+_{x_0} := \{ x \in \mathbb{R}^d : x(t) = x \text{ for some } t \in [0, T) \} \).

Given a differentiable curve \( x : [0, T) \to \mathbb{R}^d \), the arc length of \( x \) at time \( t < T \) is given by \( L_x(t) := \int_0^t \| \dot{x}(s) \| \, ds \).

**Definition 3 (Arc-Length Reparametrization).** Suppose \( x : [0, T) \to \mathbb{R}^d \) is a differentiable curve in \( \mathbb{R}^d \) with arc length at time \( t \) given by \( L_x(t) \). We say that \( \tilde{x} : [0, 1] \to \mathbb{R}^d \), \( I = [0, L_x(T)] \) is an arc-length reparametrization of \( x \) if there holds \( x(t) = \tilde{x}(L_x(t)) \) for all \( t \in [0, T) \).

As the name suggests, if \( \tilde{x} \) is an arc-length reparametrization of some curve \( x \), then the orbits of \( \tilde{x} \) and \( x \) coincide, but \( L_{\tilde{x}}(t) = t \) for all \( t \) in the maximal interval of existence.

We say that a property holds for almost every element in a set \( A \subset \mathbb{R}^d, d \geq 1 \), if the subset of \( A \) where the property fails to hold has \( L^d \)-measure zero. Likewise, we say that a property holds for almost every solution of an ODE if the property holds for solutions starting from almost every initial condition.

### III. Saddle-Point Behavior of GD and NGD: Examples and Intuition

#### A. Saddle Points and GD

The following simple example illustrates the behavior of GD near saddle points.

**Example 4.** Suppose the objective function is given by

\[
f(x) = \frac{1}{2} x^T Ax, \quad A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}
\]

and note that the origin is a saddle point of \( f \). The associated GD dynamics (1) reduce to a simple linear system of the form

\[
\frac{d}{dt} x(t) = -Ax(t)
\]

with solution \( x(t) = e^{-At} x_0 \), for initial condition \( x(0) = x_0 \in \mathbb{R}^2 \).

By classical linear systems theory, we see that solutions of this system will only converge to the origin if they start with initial conditions in the stable eigenspace of \(-A\), which is given by \( E_\sigma := \{ x = (x^1, x^2) \in \mathbb{R}^2 : x^2 = 0 \} \). Note that this is a set of initial conditions with Lebesgue measure zero.

Let \( r > 0 \) and consider the following question: What is the maximum amount of time that a solution of (4) may spend in a ball of radius \( r > 0 \) about the origin? It is straightforward to verify that trajectories not converging to 0 may take arbitrarily long to leave \( B_r(0) \), and so the time it could potentially take to escape saddle points is unbounded. Indeed, note that for \( \varepsilon \in (0, r) \), a trajectory of (4) starting on \( \partial B_\varepsilon(0) \) that enters \( B_r(0) \) must spend at least time \(-r \log(\varepsilon)\) inside the \( r \)-ball before it may enter the \( \varepsilon \) ball.

These same basic properties generalize to GD in higher dimensional systems: Solutions of GD may only converge to a saddle point from a set of initial conditions with measure zero, but the time required to escape neighborhoods of the saddle is always infinite. This is made precise in the following remark.

**Remark 5 (Saddle-Point Escape Time of GD).** Informally, given a function \( f \), a saddle point \( x^* \) of \( f \), and an \( r > 0 \) we refer to the “saddle-point escape time” of an optimization process as the maximum amount of time a trajectory that does not converge to \( x^* \) may spend in a ball of radius \( r \) about \( x^* \). In GD, the saddle point escape time is always infinite. That is, for arbitrary objective function \( f \), saddle point \( x^* \), and radius \( r > 0 \) there holds

\[
\sup_{x_0 \in \partial B_r(x^*)} \mathbb{E} \left( \left\{ t \in [0, \infty) : x_{x_0}(t) \in B_r(x^*) \right\} \right) = \infty,
\]

where \( x_{x_0} \) is the solution of (1) with initial condition \( x_0 \).
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2019.2914998, IEEE Transactions on Automatic Control

This is precisely the issue that causes GD to perform poorly in high-dimensional problems with many saddle points. In this paper we will see that NGD significantly mitigates this issue—rather than having an infinite saddle-point escape time, the saddle point escape time of NGD is at most $5\sqrt{Kn}$, where $\kappa$ is the condition number of $D^2 f(x^*)$.

B. Saddle Points and NGD

We will now consider the behavior of NGD near the saddle point in the above example.

In order to better understand this issue, it is helpful to characterize the relationship between GD and NGD. In Section IV we will see that GD and NGD are closely linked—the systems are topologically equivalent [23] and solutions of NGD are merely arc-length reparameterizations of GD solutions. In practical terms, this means that if one considers orbits of NGD and GD starting from the same initial condition $x_0 \in \mathbb{R}^d$, the orbits generated by the two systems are identical. The solutions of each system only vary in how quickly they move along the common orbit. In particular, since NGD always “moves with speed 1” (i.e., $\|x(t)\| = 1, \forall t \geq 0$) the length of an arc generated by NGD up to time $t$ is precisely $t$ (this is what it means to be an arc-length reparameterization). As an important consequence of this characterization, we will see that NGD almost never converges to saddle points (see Theorem 8).

While a solution of GD may move arbitrarily slowly as it passes near a saddle point, a solution of NGD starting at the same initial condition will move along the same orbit with constant speed, not slowing near the saddle point. This is illustrated in Fig. 1.

Consider NGD with $f$ as defined in Example 4 (see (3)). Given the simple linear structure of the corresponding GD ODE (4), it is straightforward to verify that the arc length of any trajectory of GD (or equivalently NGD) intersecting $B_r(0)$ is upper bounded by $2r$ and hence the maximum time a trajectory of NGD may spend in $B_r(0)$ is $2r$ (see Fig. 1).

This simple example may be generalized to higher dimensions. Let $f : \mathbb{R}^d \to \mathbb{R}^d, d \geq 2$ be given by $f(x) = x^T Ax$, with $A = \text{diag}(\lambda_1, \ldots, \lambda_d), |\lambda_i| = 1$ for all $i = 1, \ldots, d$, and at least one $\lambda_i > 0$ and one $\lambda_i < 0$. Given the simple linear structure of the corresponding GD ODE $\dot{x} = -Ax$, it is straightforward to show that the arc-length of any trajectory of GD intersecting $B_r(0)$ (and hence the amount of time spent by NGD in $B_r(0)$) is upper bounded by $2r$, independent of the dimension $d$.

Note that in this example, the condition number of $D^2 f(0)$ is 1. In general, as the condition number increases, the time spent by NGD in $B_r(0)$ may increase. Theorem 9 captures this relationship for general $f$ (satisfying Assumption 1).

Remark 6. We note that the bound to be established in Theorem 9 is conservative. In particular, suppose $f : \mathbb{R}^d \to \mathbb{R}$ is quadratic of the form $f(x) = x^T Ax$, with $A \in \mathbb{R}^{d \times d}$ diagonal and non-singular. Then one can show that time spent by a trajectory of NGD in $B_r(0)$ is at most $2\sqrt{dr}$. This bound holds even as the condition number of $D^2 f(0)$ is brought to $\infty$.

Thus, while an ill-conditioned saddle point can slow the escape time of NGD, this example suggests that in the worst case, as the condition number is brought to $\infty$, the time spent by NGD in $B_r(x^*)$ about a saddle point $x^*$ can be bounded by $C\sqrt{dr}$, where $C > 0$ is some universal constant independent of dimension and condition number. An in-depth investigation of this issue is outside the scope of this paper.

IV. NGD: STRUCTURAL PROPERTIES AND GENERIC CONVERGENCE TO LOCAL MINIMA

The following proposition establishes the basic structural relationship between GD and NGD.

Proposition 7. Let $x(t)$ and $\dot{x}(t)$ be solutions of (1) and (2) respectively, with the same initial condition $x_0$, over maximal intervals of existence $[0,T)$ and $[0,T)$, respectively. Then $\dot{x}(t)$ is an arc-length reparametrization of $x(t)$, and $\dot{x}(t) = x(s(t))$ for some strictly increasing function $s(t)$, with $s(0) = 0$ and $s(\tilde{T}) = T$.

We recall that the notion of an arc-length reparametrization can be found in Definition 3.

This result means that (classical) solutions of (1) and (2) starting at the same initial condition have identical orbits; the solutions only differ in the speed with which they move along the common orbit.5

The following result shows that NGD may only converge to non-degenerate saddle points from a measure-zero set of initial conditions. The proposition considers a slightly weaker condition than non-degeneracy as discussed in Section II. In particular, we will require that at least one eigenvalue of $D^2 f(x^*)$ be negative. Saddle points satisfying this condition

4This is shown by bounding the arc length of the corresponding linear GD ODE $\dot{x} = -Ax$. Intuitively, if $A$ is well conditioned, then trajectories of the ODE passing near 0 travel along a “direct route” to and away from 0. If $A$ is ill conditioned, then trajectories of the ODE travel a “Manhattan route” to and away from 0, with movement tangential to the stable eigenspace of $A$ occurring along only one stable eigenvector at a time.

5In other words, the dynamical systems defined by (1) and (2) are topologically equivalent [23] with the concomitant homeomorphism given by the identity.
are sometimes referred to in the literature as strict (or rideable) saddle points \[6\], \[24\].

**Theorem 8** (Non-Convergence to Saddle Points).

(i) Let \( x^* \) be a saddle point of \( f \) such that there exists a \( \lambda \in \sigma(D^2 f(x^*)) \) with \( \lambda < 0 \). Then solutions to (2) can only reach or converge to \( x^* \) from a set of initial conditions with Lebesgue measure zero.

(ii) Suppose that each saddle point of \( f \) is non-degenerate. Then the set of initial conditions from which solutions to (2) reach or converge to a saddle point has Lebesgue measure zero.

Since a non-degenerate saddle point \( x^* \) necessarily has at least one strictly negative eigenvalue in \( \sigma(D^2 f(x^*)) \), Theorem 8 immediately implies that solutions to (2) may only converge to non-degenerate saddle points from a measure-zero set of initial conditions.

It follows from Propositions 7 and 8 that solutions of NGD exist and are unique for almost every initial condition. We note that both of these results follow as elementary applications of classical ODE theory (see Section VII).

We also note that this issue (generic non-convergence to saddle points, as in Proposition 8) was considered for discrete-time GD in the recent work [7].

\section{Fast Escape From Saddle Points}

The following theorem gives our main result regarding fast escape from saddle points. The theorem provides a simple upper bound on the amount of time that trajectories of NGD can spend near saddle points.

**Theorem 9** (Saddle-Point Escape Time). Let \( C > 4 \) and suppose \( x^* \) is a non-degenerate saddle point of \( f \). Then for all \( r > 0 \) sufficiently small, any trajectory of (2) that does not reach or converge to \( x^* \) can spend time at most time \( C\sqrt{kr} \) in the ball \( B_r(x^*) \), where \( k \) is the condition number of \( D^2 f(x^*) \).

That is, if \( x_{x_0} \) is a solution to (2) with initial condition \( x_0 \) and maximal interval of existence \([0,T_{x_0}]\), then

\[
L^1\left\{ t \in [0,T_{x_0}] : x_{x_0}(t) \in B_r(x^*) \right\} \leq C\sqrt{kr}.
\]

We recall that by Theorem 8, solutions of (2) can only reach saddle points from a set of initial conditions with measure zero, hence the theorem holds for solutions starting from almost every initial condition.

In order to underscore the significance of this result, we recall that the saddle-point escape time of GD (i.e., the time required to escape a ball of radius \( r > 0 \) about a saddle point) is infinite, independent of \( f, d, x^* \), and \( r \) (see Remark 5). This is precisely the issue that causes slow convergence of GD in problems with many saddle points. In contrast to this, Theorem 9 shows that trajectories of NGD always escape a ball of radius \( r \) within time \( 5\sqrt{kr} \).\(^6\)

Furthermore, we recall that Proposition 7 showed that orbits of GD and NGD coincide. Thus, away from saddle points (where GD is generally “well behaved”), GD and NGD behave in an essentially identical manner.

A few remarks are now in order.

**Remark 10** (Values of constant \( C \)). The above theorem holds with the constant \( C \) set to any value strictly greater than 4. The proof of the estimate in the theorem utilizes several Taylor series approximations. There is a tradeoff inherent in this proof technique—as \( C \) approaches 4, the range of permissible values of \( r > 0 \) where the Taylor approximation (and hence, the theorem) is applicable shrinks to zero. For clarity of presentation and to emphasize the key features of this result we find it convenient to simply fix the constant to be 5 in the abstract and introduction. See Proposition 19 and proof thereof for more details.

**Remark 11** (Permissible values of \( r \)). The range of values of \( r > 0 \) where Theorem 9 holds depends both on the constant \( C \) and the magnitude of higher order derivatives near the saddle point \( x^* \). In particular, the result holds so long as the Taylor estimates (8), (9) used in the proof are valid. If one assumes that \( f \) is more than twice differentiable and assumes bounds on the magnitude of the higher order derivatives near \( x^* \), then the radius where these estimates hold can be bounded, and a more precise statement can be made about the permissible values of \( r \) in Theorem 9. For example, if one assumes that \( |D^3 f(x)| < \hat{C} \) is uniformly bounded for some \( \hat{C} > 0 \) then Theorem 9 holds for all \( r \in (0, \tilde{r}) \), where

\[
\tilde{r} = 6\kappa^{-1/2}\hat{C}^{-1}\left|\lambda_{\max}(D^2 f(x^*))\right|^{\frac{2\kappa}{\tilde{r}}}
\]

and where \( \kappa \) is the condition number of \( D^2 f(x^*) \). This may be verified by noting that the Taylor estimates (8), (9) used in the proof of Proposition 19 are valid in the ball \( B_{\tilde{r}}(0) \), \( \tilde{r} = \kappa^{1/2}r \), for values of \( r \) in this range.

**Remark 12** (Non-Applicability of the Hartman-Grobman Theorem). The Hartman-Grobman theorem from classical differential equations states that near non-degenerate saddle points one can construct a homeomorphism mapping the trajectories of a non-linear ODE to trajectories of the associated linearized system [23]. It is simple to show that Theorem 9 holds when \( f \) is quadratic (and hence the associated NGD system is topologically identical to a linear system); see Section III-A. Thus, one might expect Theorem 9 to hold for general (non-quadratic) \( f \) by the Hartman-Grobman theorem. However, the homeomorphisms constructed in the Hartman-Grobman theorem are in general not smooth, and so will not preserve trajectory length, and cannot be used to prove a bound such as Theorem 9. Instead one must resort to more analytical techniques to study path length; see the proof of Proposition 19 below.

**Remark 13** (Theorem 9 Proof Technique). The key idea of the proof of Theorem 9 relies on establishing a differential inequality between the “potential” \( f \) and the “potential dissipation rate” \( \frac{d}{dt} f(x(t)) \). The methods are flexible, and may be applicable to other non-smooth settings. In a previous work [25] the authors utilized similar techniques to study non-smooth dynamics in game-theoretical problems.
VI. A GLOBAL CONVERGENCE-TIME BOUND

We will now use the above results to prove a simple corollary bounding the maximum amount of time that trajectories can take to converge to a local minimum under (2). We do not claim that these results are optimal; we present them as a simple application of the tools developed in this work.

To achieve this goal, we make modest assumptions (which have been utilized in similar settings previously [6], [16]) that guarantee $D^2 f$ is uniformly non-degenerate at critical points, and that the slope of $f$ is bounded away from zero away from critical points. These assumptions permit us to bound the amount of time spent both near saddle points (via Theorem 9) and the amount of time spent away from critical points (via an assumption that $\nabla f$ is uniformly bounded away from zero away from critical points). More precisely, we make the following assumptions:

**Assumption 14.** The function $f$ is of class $C^3$ and $|D^3 f(x)| \leq \hat{C}$ uniformly for all $x \in \mathbb{R}^d$, for some $\hat{C} > 0$.

This assumption ensures that there exists a single $\bar{r} > 0$ such that Theorem 9 (and Proposition 19) hold uniformly for any $r \in (0, \bar{r})$ and at every critical point. The fact that such an $\bar{r}$ is guaranteed to exist under Assumption 14 is discussed in more detail in Remark 11.

Next we assume a uniform bound on the magnitude of eigenvalues of the Hessian at critical points.

**Assumption 15.** There exist constants $|\lambda|_{\max}, |\lambda|_{\min} > 0$ such that for every critical point $x^*$ of $f$ there holds $|\lambda|_{\min} \leq |\lambda| \leq |\lambda|_{\max}$ for all $\lambda \in \sigma(D^2 f(x^*))$.

The next assumption ensures that at any point $x \in \mathbb{R}^d$, either the gradient of $f$ at $x$ is large (guaranteeing fast local improvement of descent techniques), or $x$ is close to a critical point.

**Assumption 16.** Fix $C > 4$. Assuming Assumptions 14 and 15 hold, let $r > 0$ be chosen so that Theorem 9 (or equivalently, Proposition 19) holds with constant $C$ at every critical point and so that $r \leq \frac{|\lambda|_{\min}}{C}$. Furthermore, assume that there exists a constant $\nu > 0$ such that for all $x \in \mathbb{R}^d$ either

$$\|x - x^*\| < r, \text{ with } \nabla f(x^*) = 0 \text{ or } \|\nabla f(x)\| > \nu.$$ 

Assumptions 15 and 16 together are similar to the strict saddle property assumed in [6], [16]. The main difference is that here we assume a uniform (lower) bound on the minimum-magnitude eigenvalue of the Hessian at all critical points rather than just saddle points, and we assume a uniform (upper) bound on the maximum-magnitude eigenvalue of the Hessian at all critical points.

The final assumption ensures that a descent process will eventually converge to some point rather than expanding out infinitely. This assumption is naturally satisfied if, for example, $f$ is coercive (i.e., $f(x) \to \infty$ as $\|x\| \to \infty$).

**Assumption 17.** There exists an $R > 0$ such that trajectories of (2) that begin in $B_R(0)$ remain in $B_R(0)$ for all $t \geq 0$.

Let $R > 0$ be as in Assumption 17 and let

$$M := \sup_{x \in B_R(0)} |f(x)|. \quad (5)$$

Note that since $f$ is continuous, $M < \infty$.

The following result gives a simple estimate on the amount of time the dynamics (2) will take to reach a local minimum.

**Corollary 18.** Suppose that every saddle point $x^*$ of $f$ is non-degenerate and that Assumptions 14–17 hold. Then for almost every initial condition inside $B_R(0)$, solutions of (2) will converge to a local minima in at most time $2M\nu^{-1} + C\sqrt{\frac{|\lambda|_{\max} \|r + r\|}{\|r\|}}$, where $C > 4$ is the constant in Assumption 16.

We remark that the estimate obtained in the above result can be conservative. In particular, Theorem 9 (and Proposition 19) are only shown to hold for $r$ sufficiently small. For small values of $r$, it can occur that the estimate $\|\nabla f(x)\| > \nu$ in Assumption 16 is conservative, which (together with Remark 6) can lead to very conservative time estimates in Corollary 18.

VII. PROOFS OF MAIN RESULTS

We now present the proofs of the results found in Sections IV – V.

We begin by presenting the proofs of Propositions 7 and Theorem 8, which follow from elementary applications of classical ODE theory.

**Proof of Proposition 7.** Given a solution $x$ (with arc length at time $t$ given by $L_x(t)$) to (1), one can reparametrize the trajectory by arc length, i.e., $\hat{x}(L_x(t)) = x(t)$, and $\frac{d}{dt} \hat{x}(t) = -\frac{\hat{x}(t)}{\|\nabla \hat{x}(t)\|}$. Using the chain rule we find that $\frac{d}{dt} \hat{x}(t) = -\frac{\hat{x}(t)}{\|\nabla \hat{x}(t)\|}$. Since the solutions are classical, uniqueness of solutions for ODEs gives us that $\hat{x}$ and $\hat{x}$ must be equal.

**Proof of Theorem 8.** We begin by proving part (i) of the theorem. Solutions to (1) which converge to some saddle point are contained within a stable manifold, i.e., a smooth surface of at most dimension $n - 1$. Such a surface will be a set with Lebesgue measure zero. Proof and details of such a result may be found in [23]. The result then follows from Proposition 7.

Part (ii) of the theorem follows from the fact that if all saddle points are non-degenerate, then all saddle points are isolated. Hence, the set of saddle points is countable. By part (i) of the theorem, the union of the stable manifolds for all saddle points is a set with Lebesgue measure zero.

The following proposition proves Theorem 9. The proposition is stated in slightly more general terms than Theorem 9 in order to account for the behavior of NGD near minima as well as saddle points.

**Proposition 19.** Let $C > 4$, let $x^* \in \mathbb{R}^d$ be a non-degenerate critical point of $f$, and let $x(t)$ be a solution of (2) with arbitrary initial condition $x_0 \neq x^*$. For all $r > 0$ sufficiently small, the time spent in $B_r(x^*) \setminus \{x^*\}$ is bounded according to

$$\mathcal{L}^1\left\{t \geq 0 : x(t) \in B_r(x^*) \setminus \{x^*\}\right\} \leq C\sqrt{\kappa r},$$
where $\kappa = \frac{|\lambda_{\max}(D^2f(x^*))|}{|\lambda_{\min}(D^2f(x^*))|}$.

**Proof.** Without loss of generality, assume $x^* = 0$ and let $H := D^2f(0)$. For $x \in \mathbb{R}^d$ define $\tilde{d}(x) := \sqrt{x^THx}$, where $|B| := \sqrt{BB^T}$ for a square matrix $B$. The function $\tilde{d}$ will be a convenient modified distance for the proof. For convenience in notation, throughout the proof we use the shorthand $|\lambda_{\max}| := |\lambda_{\max}(D^2f(x^*))|$ and $|\lambda_{\min}| := |\lambda_{\min}(D^2f(x^*))|.$

Note that for $a \geq 0$ we have the following relationships

$$\|x\| \leq \frac{a}{\sqrt{|\lambda_{\min}|}} \implies \tilde{d}(x) \leq a \quad (6)$$

$$\tilde{d}(x) \leq a \implies \|x\| \leq \frac{a}{\sqrt{|\lambda_{\min}|}}. \quad (7)$$

By Taylor’s theorem and the non-degeneracy of $x^*$, for any $C_1 > \frac{1}{2}$ there exists a neighborhood of 0 such that

$$|f(x) - f(0)| \leq C_1 d(x)^2. \quad (8)$$

Using the chain rule we see that along the path $x(t)$, the potential changes as

$$\frac{d}{dt} f(x(t)) = -\|\nabla f(x(t))\|.$$

Let $C_2 < 1$ be arbitrary. Again using Taylor’s theorem and the non-degeneracy of $H$, for $x(t)$ in a neighborhood of 0 we have that

$$\|\nabla f(x(t))\| \geq C_2 \|Hx(t)\| \geq C_2 \|H^{1/2}H^{1/2}x(t)\| \geq C_2 \sqrt{|\lambda_{\min}| \|H^{1/2}x(t)\|} = C_2 \sqrt{|\lambda_{\min}|} \tilde{d}(x(t)). \quad (9)$$

In turn

$$-\frac{d}{dt} f(x(t)) \geq C_2 \sqrt{|\lambda_{\min}|} \tilde{d}(x(t)). \quad (10)$$

Let $\hat{r} > 0$ be such that the estimates (8) and (9) hold inside the closed ball $B_{\hat{r}}(0)$. Suppose that $x(t) \in B_{\hat{r}}(0)$ for $t \in [t_1, t_2]$. Letting $\varepsilon(t) := \tilde{d}(x(t))$ and integrating (10) gives

$$f(x(t_1)) - f(x(t_2)) \geq C_2 \sqrt{|\lambda_{\min}|} \int_{t_1}^{t_2} \varepsilon(s)ds. \quad (11)$$

Let $r := \kappa^{-1/2}\hat{r}$ and $\hat{r}$ is the radius of the ball where (8) and (9) hold and is dependent on $C_1$ and $C_2$. By choosing $C_1$ close to $\frac{1}{2}$ and $C_2$ close to 1, the constant $C$ may be brought arbitrarily close to 4 with the range of permissible values of $r$ changing accordingly. This proves the desired result. \[\square\]

**Proof of Corollary 18.** First, we claim that critical points must be separated by a distance of at least $2r$. Let $x^*$ be a critical point. Then

$$\nabla f(x) = \int_{0}^{1} D^2f((1-s)x^* + sx)(x-x^*) \, ds$$

$$= \int_{0}^{1} D^2f(x^*)(x-x^*) \, ds$$

$$+ \int_{0}^{1} \int_{0}^{1} D^3f((1-\tau)x^* + \tau x)(x-x^*) \, d\tau \, ds,$$

where by $D^3f_{-x}$ we mean the matrix representing the third derivative evaluated in the direction $x - x^*$. We can then bound

$$|\nabla f(x)| \geq |\lambda_{\min}| |x-x^*| - \frac{C}{2} |x-x^*|^2$$

where $C$ is the bound on our third derivatives. Note that by Assumption 16 we have $2r \leq \frac{2|\lambda_{\max}|}{C}$. Thus we see that for any $x \in B_{2r}(x^*) \{x^*\} \subset B_{\frac{2|\lambda_{\max}|}{C}}(x^*)$ we have $\nabla f(x) \neq 0$. Subtracting $f(0)$ to the left-hand side above and using (8) we obtain

$$\frac{2C_1}{C_2 \sqrt{|\lambda_{\min}|}} \varepsilon^2 \geq \int_{\varepsilon(s) \leq \eta} e(s)ds.$$

Markov’s inequality [26] then gives

$$L^1\left(\{s : \eta \geq \varepsilon(s) \geq \frac{\eta}{2}\}\right) \leq \frac{2}{\eta} \int_{\varepsilon(s) \leq \eta} e(s)ds \leq \frac{4C_1}{\eta C_2 \sqrt{|\lambda_{\min}|}} \varepsilon^2 \leq \frac{1}{8C_1} \eta.$$
Hence critical points must be separated by a distance of at least $2r$.

Now, let $x(t)$ be a classical solution of (2) (which, by Theorem 8, holds for a.e. solution of (2)). Let $[t_1, t_2] = I$ be the maximal interval of existence for this classical solution. Our goal is to prove that $(t_2 - t_1)$ can be covered uniformly.

To this end, we divide $I$ into two subsets, $I_c, I_o$, where $I_c$ are the times where $\|x(t) - x^*\| \leq r$ for some critical point $x^*$, and $I_o$ are points where $\|\nabla f(x(t))\| \geq \nu$.

Using the chain rule we see that $\frac{d}{dt} f(x(t)) = -\|\nabla f(x(t))\|$. By Assumption 17 and (5) we have $|f(x(t))| < M$ along any trajectory of (2) starting in $B_R(0)$. Thus, we immediately have that $|I_o| < 2M\nu^{-1}$, where we use the notation $|\cdot|$ to denote the length of the interval.

Let $\kappa = \frac{|I_o|}{|I_c|}$. By Proposition 19 we can spend at most $C\sqrt{\kappa r^2}$ near any particular critical point. Since critical points are separated by at least distance $2r$, we can cover all the critical points with disjoint balls of radius $r$. By then estimating the volume, the total number of critical points within distance $R$ of the origin is at most $\frac{(R+r)^d}{d\pi^{d/2}}$. Thus, we find that $|I_o| < C\sqrt{\kappa r^2} \frac{(R+r)^d}{d\pi^{d/2}}$.

In summary, we find that $|I| \leq 2M\nu^{-1} + C\sqrt{\kappa r^2} \frac{(R+r)^d}{d\pi^{d/2}}$. This implies that classical trajectories can be of length at most $M\nu^{-1} + C\sqrt{\kappa r^2} \frac{(R+r)^d}{d\pi^{d/2}}$. Since a.e. initial condition does not reach any saddle point, almost every initial condition will converge to a local minimizer of $f$ in $2M\nu^{-1} + C\sqrt{\kappa r^2} \frac{(R+r)^d}{d\pi^{d/2}}$ time. This concludes the proof.

VIII. CONCLUSION

The paper considered the properties of NGD near saddle points. It was shown that NGD may only converge to a saddle point from a low-dimensional stable manifold and that NGD escapes quickly from the neighborhood of saddle points. As a simple example application of these results, a global bound on the convergence time of NGD was established. An important future research direction is to rigorously establish analogous stable-manifold results and escape-time bounds for discrete-time NGD. A shortcoming of the present work was noted in Remark 6; namely, the bound established in Theorem 9 can be quite conservative when the condition number $\kappa$ is large. Future work may consider establishing a dimension-dependent (and condition number independent) bound on saddle-point escape time as suggested in Remark 6.

REFERENCES