SIAM J. MATH. DATA SCI.
Vol. 2, No. 4, pp. 1181–1197

Extending the Step-Size Restriction for Gradient Descent to Avoid Strict Saddle Points*

Hayden Schaeffer† and Scott G. McCalla‡

Abstract. We provide larger step-size restrictions for which gradient descent-based algorithms (almost surely) avoid strict saddle points. In particular, consider a twice differentiable (nonconvex) objective function whose gradient has Lipschitz constant L and that the set of points that obtain the maximum value of the spectral norm of the Hessian is measure zero. We prove that given one uniformly random initialization, the probability that gradient descent with a step-size up to $2/L$ will converge to a strict saddle point is zero. This extends previous results up to the sharp limit imposed by the convex quadratic case (provably converging to local minimizers). In addition, the arguments hold in the case when a learning rate schedule is given, with either a continuous decaying rate or a piecewise constant schedule. We show that the assumptions are robust in the sense that functions which do not satisfy the assumptions are meager with respect to analytic functions.

Key words. gradient descent, nonconvex, strict saddles, optimization

AMS subject classifications. 90C26, 37L10

DOI. 10.1137/19M129509X

1. Introduction.

Gradient descent-based methods are among the main algorithms for optimizing models throughout machine learning. As many learning models are nonconvex, their energy landscapes may consist of spurious local minima and saddles; this may lead algorithms to learn models that do not generalize well to new data [12]. In [22], it was argued that in high-dimensional optimization, saddle points are more problematic than local minima. It is easy to construct examples for which gradient descent converges to saddle points given certain initializations [18, 14]. However, when the step-size is sufficiently small and the saddles are *strict*, i.e., the Hessian has at least one negative eigenvalue, the gradient descent method is unlikely to converge to a saddle [14]. On the other hand, it is still possible that gradient descent will take exponential time to escape [7]. The strict saddle condition appears in many applications, for example, orthogonal tensor decomposition [8], low-rank matrix recovery [3, 9, 10], dictionary learning [28, 29], generalized phase retrieval [30], and neural networks [27].

First-order gradient descent-based methods can avoid or escape saddles when unbiased noise is added to the system. In [23], the authors prove that the Robbins–Monro stochastic

*Received by the editors October 23, 2019; accepted for publication (in revised form) September 25, 2020; published electronically December 18, 2020.
https://doi.org/10.1137/19M129509X
Funding: The first author acknowledges the support of AFOSR, FA9550-17-1-0125 and NSF CAREER grant 1752116. The second author acknowledges the support of NSF grant 1813654 and the Army Research Office (W911NF-19-1-0288).

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 USA (hschaeff@andrew.cmu.edu).
‡Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717 USA (scott.mccalla@montana.edu).

approximation converges to local minima in the presence of strict saddles. For objective functions with strict saddles, [8] provided quantitative convergence rates to local minima for the noisy gradient descent method. Convergence of the normalized gradient descent with noise injection was shown in [15].

Alternatively, deterministic methods which use second-order information or trust regions [5] have a rich history and can be used to avoid strict saddles. Some examples of such methods include the modified Newton's method using negative curvature [17], the cubic-regularized Newton's method [19], the saddle-free Newton's method for deep learning [6, 22], algorithms for higher-order saddles [2], and trust-region approaches in [28, 29, 30].

One issue with "second-order" approaches is the need for higher-order information that leads to polynomial (in dimension) complexity per iteration. For machine learning problems, which are typically of very high dimension, this complexity can be prohibitive. Some recent approaches [24, 25, 4] were proposed to lower the per-iteration complexity of second-order methods while converging to second-order stationary points (see [4]). In [11], the authors propose a perturbed gradient descent method which converges to the second-order condition with a poly-logarithmic cost.

Contributions of this work. The recent work of [14, 21, 13] showed that, under various conditions, the gradient descent algorithm will avoid strict saddle points (without the need for additional hyper-parameters or higher-order information). The main technique is to show that the attracting set of a strict saddle has zero measure by invoking the stable manifold theorem applied to the discrete dynamical system generated by the gradient descent method for a C^2 nonconvex objective function f with time-step $\alpha > 0$. In [14], it was proved that gradient descent avoids strict saddles if the gradient of the objective function has Lipschitz constant L (globally), isolated saddle points, and $\alpha < 1/L$. In [13], it was shown that many first-order methods will avoid strict saddles under these conditions. Accelerated methods, such as the heavy-ball method, also avoid strict saddles, as shown in [20].

The results still hold with weaker conditions. In particular, [21] showed that a nonglobal Lipschitz constant L (in a convex forward invariant set) and $\alpha < 1/L$ were enough. If the objective function is coercive, then the sublevel sets are compact and L does not have to be global; however, the results of [21] hold more generally. They also showed that over the set of all local minimizers C, if

$$0 < \gamma < \inf_{x \in C} ||\nabla^2 f(x)||_2 < \infty,$$

then $\alpha < 2/\gamma$ is a necessary condition for gradient descent to converge to a local minimizer.

There are still several open questions, in particular, if the time-step restriction $\alpha < 1/L$ is necessary for avoiding strict saddles and if varying time-steps affects these results [14, 13]. In this work, we show that if the set of points that obtain the maximum value of the spectral norm of the Hessian is measure zero, then the bound can be extended to $\alpha = 1/L$. Furthermore, a time-step of $\alpha < 2/L$ is possible if α^{-1} is not equal to an eigenvalue of the Hessian outside of a null set. Examples highlight the need for such conditions. In addition, we show that these arguments can apply to gradient descent with a varying time-steps.

2. Overview and examples.

To solve the nonconvex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

consider the gradient descent method with fixed step-size $\alpha > 0$, i.e.,

$$x^{n+1} = x^n - \alpha \nabla f(x^n).$$

The sequence x^n is generated by the iterative map $x^{n+1} = g(x^n) = g^n(x^0)$, where $g^n = g \circ g^{n-1}$ and

$$g(x) := x - \alpha \nabla f(x).$$

Given conditions on f and α, the method will converge to a critical point of f (or equivalently a fixed point of the map g) [1]. Let $\lambda_j(\cdot)$ denote the jth eigenvalue of a matrix (in descending order).

Definition 2.1. *Consider a function $f : \mathbb{R}^d \to \mathbb{R}$, and assume $f \in C^2(\mathbb{R}^d)$. We define the following:*

- *A point $x^* \in \mathbb{R}^d$ is a critical point of f if $\nabla f(x^*) = 0$.*
- *A critical point $x^* \in \mathbb{R}^d$ is a strict saddle point if there is a negative eigenvalue, i.e.,*

$$\lambda_j\left(\nabla^2 f(x^*)\right) < 0$$

for some $1 \le j \le d$.

Based on this definition, some local maxima are technically strict saddle points. Saddle points like $(0,0)$ of the objective function $x^2 - y^3$ are avoided by the definition of strict saddles.

Define L as the smallest Lipschitz constant of the gradient over a domain Ω. If $f \in C^2(\Omega)$, then it is easy to see that

$$L := \sup_{x \in \Omega} \|\nabla^2 f(x)\|_2.$$

It was shown in [14, 21, 13] that for $\alpha < L^{-1}$, gradient descent avoids strict saddle points. Extending this result to $\alpha \le L^{-1}$ introduces issues even for smooth objective functions. It is possible for gradient descent to converge to strict saddles if there are nontrivial regions where g degenerates (i.e., the Jacobian Dg is noninvertible). In effect, the gradient flow funnels iterates toward the stable manifold of a strict saddle. To illustrate various issues, we present the following examples.

Example 2.2 (from [18, 14]). Consider the objective function

$$f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$$

over $\Omega = \mathbb{R} \times (-\sqrt{\frac{11}{3}}, \sqrt{\frac{11}{3}})$, which has three critical points $(0,0)$ (strict saddle) and $(0, \pm 1)$ (minima). The Hessian is given by

$$\nabla^2 f(x,y) = \begin{bmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{bmatrix},$$

and the value of the spectral norm is maximized over Ω when $y = \pm\sqrt{\frac{11}{3}}$, i.e.,

$$L = \sup_{\Omega} \|\nabla^2 f(x,y)\| = 10.$$

The gradient descent method with step-size $\alpha = L^{-1} = \frac{1}{10}$ is given by

$$\begin{bmatrix} x^{n+1} \\ y^{n+1} \end{bmatrix} = \begin{bmatrix} \frac{9}{10}x^n \\ \frac{11}{10}y^n - \frac{1}{10}(y^n)^3 \end{bmatrix}.$$

The system is forward invariant over Ω, with x^n converging to 0 and y^n converging to $\text{sign}(y^0)$ (minima). The sequence will only converge to the strict saddle point $(0,0)$ on the line $(x,0)$ and thus has probability zero if the initial data are sampled uniformly from Ω.

Example 2.3. Consider the objective function

$$f(x,y) := \frac{1}{4}y^2 - q(y)x^2$$

for some region of $\mathbb{R}^2$ containing the origin, and let $q \in C^2$. The gradient is given by

$$\nabla f(x,y) = \begin{bmatrix} -2q(y)x \\ \frac{1}{2}y - q'(y)x^2 \end{bmatrix},$$

and the Hessian is given by

$$\nabla^2 f(x,y) = \begin{bmatrix} -2q(y) & -2q'(y)x \\ -2q'(y)x & \frac{1}{2} - q''(y)x^2 \end{bmatrix}.$$

If we define q as a smooth interpolant between 1 and -1 for $y \in (10,30)$, then we can show that even though the critical point at $(0,0)$ is a strict saddle and the flow is invertible near the strict saddle, regions of degeneracy away from the strict saddle can converge to the stable manifold and thus with some nonzero probability converge to a strict saddle.

For an explicit example, define q by

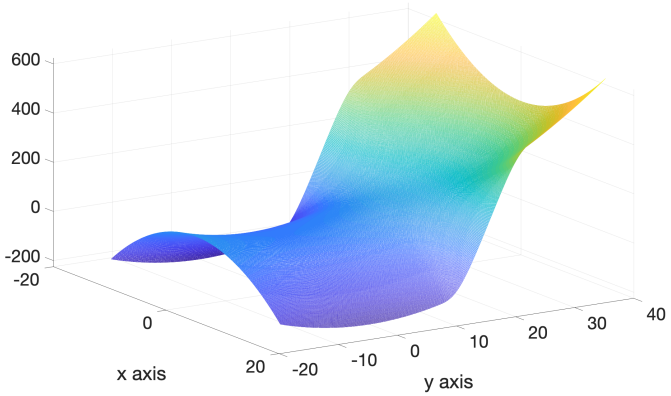
$$q(y) := \begin{cases} 1 & \text{if } \ y \le 10, \\ 1 - \frac{2}{1+\exp\left(\frac{40(y-20)}{(y-20)^2-100}\right)} & \text{if } \ y \in (10,30), \\ -1 & \text{if } \ y \ge 30. \end{cases}$$

It is easy to check that the function $q \in C^2$. In the region $y < 10$, we have

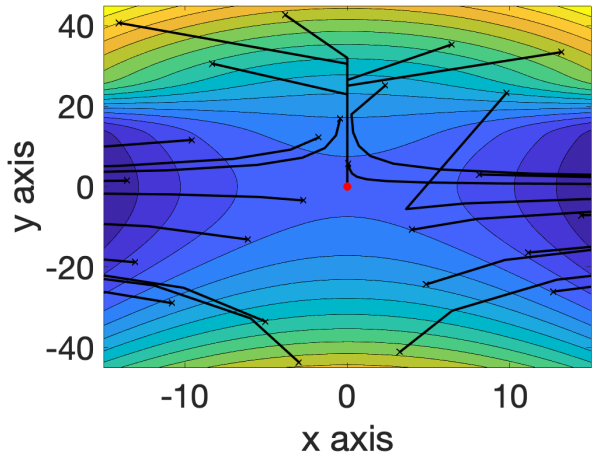
$$\nabla^2 f(x,y) = \begin{bmatrix} -2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \text{ and in the region } y > 30\text{: } \nabla^2 f(x,y) = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

The only critical point is at $(x,y) = (0,0)$, and it is a strict saddle. Note that the Lipschitz constant of ∇f in some bounded region around the strict saddle that contains $y \ge 30$, restricted to x near the origin, is given by $L = 2$ and is obtained for all $y \ge 30$ (a set of positive measure). Using gradient descent with $\alpha = L^{-1} = 1/2$ yields

$$\begin{bmatrix} x^{n+1} \\ y^{n+1} \end{bmatrix} = \begin{bmatrix} x^n + q(y^n)x^n \\ y^n - \frac{1}{4}y^n + \frac{1}{2}q'(y^n)(x^n)^2 \end{bmatrix}.$$



(a) Surface Plot



(b) Random Gradient Descent Trajectories

Figure 2.1. *The plots correspond to the surface* (a) *of the objective function from Example* 2.4 *and the associated gradient descent trajectories* (b) *from a set of randomly sampled initial conditions. The red dot at* $(0,0)$ *is the strict saddle point. The trajectories whose initial* y *values are large enough will converge to the saddle point.*

For any initialization in $y \geq 30$, we have

$$\begin{bmatrix} x^{n+1} \\ y^{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{3}{4} y^n \end{bmatrix},$$

which is within the stable manifold for $(0,0)$ (the iterates are pushed onto the stable manifold after one step). Therefore, given this choice of step-size, with nonzero probability (after restricting onto an appropriate bounded set), gradient descent will converge to a strict saddle.

In Figure 2.1, we set $\alpha = L^{-1}$ and plot the gradient descent trajectories, sampled from a set of random initial states. The figure shows that the trajectories whose initial y value are large enough are all funneled to the stable manifold of the saddle at $(0,0)$.

Example 2.3 shows that large regions of space can be attracted to the local stable manifold of a strict saddle. These domains act as focusing regions, in particular, subsets where the Hessian is degenerate (i.e., at least one zero eigenvalue) can cause the flow to focus a nonzero measure set onto a measure zero stable manifold. This behavior will be taken into account in Theorem 3.2.

In the next section, we provide qualitative and quantitative results on the convergence of gradient descent, in particular, the divergence from strict saddles when the time-step does not degenerate the Jacobian of g on nonnull sets.

3. Conditions for avoiding strict saddles.

For convex optimization problems with Lipschitz gradients, convergence of the gradient descent method is guaranteed for step-sizes satisfying $\alpha L \leq 1$. It is possible to take larger step-sizes. For example, if A is a symmetric positive definite matrix, then gradient descent with fixed step-size will converge to a minimizer of

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

if and only if $\alpha L < 2$. Taking $\alpha L < 2$ as a reasonable upper limit for the step-size, our goal is to show that with the time-step restriction and a condition on the size of the degenerate set, gradient descent will not converge to a strict saddle. Note that this does not imply convergence to a minimizer since nonstrict saddles are possible.

The behavior near a critical point can be characterized by the well-known center manifold theorem.

Theorem 3.1 (center manifold theorem [26]). *Let x^* be a fixed point of a C^1 local diffeomorphism $g : U \to \mathbb{R}^d$, where U is a neighborhood of x^* in $\mathbb{R}^d$ equipped with the Euclidean metric $d(\cdot,\cdot)$. Let $E^s \bigoplus E^c \bigoplus E^u$ be an invariant splitting of $\mathbb{R}^d$ into the generalized eigenspace of the Jacobian $Dg(x^*)$ corresponding to the eigenvalues of absolute value less than one, equal to one, and greater than one. Then for each of the invariant subspaces E^s, $E^s \bigoplus E^c$, E^c, $E^c \bigoplus E^u$, and E^u, there is an associated local g invariant C^1 embedded disc W^s_{loc}, W^{cs}_{loc}, W^c_{loc}, W^{cu}_{loc}, and W^u_{loc} tangent to the linear subspace at x^* and a ball B around x^* such that there is a norm with the following:*

1. *$W^s_{loc} = \{x \in B : g^n(x) \in B \text{ for all } n \geq 0 \text{ and } d(g^n(x), 0) \to 0 \text{ exponentially}\}$. Also, $g : W^s_{loc} \to W^s_{loc}$ is a contraction map.*
2. *$g(W^{cs}_{loc}) \cap B \subset W^{cs}_{loc}$. If $g^n(x) \in B$ for all $n \geq 0$, then $x \in W^{cs}_{loc}$.*
3. *$g(W^c_{loc}) \cap B \subset W^c_{loc}$. If $g^n(x) \in B$ for all $n \in \mathbb{Z}$, then $x \in W^c_{loc}$.*
4. *$g(W^{cu}_{loc}) \cap B \subset W^{cu}_{loc}$. If $g^n(x) \in B$ for all $n \leq 0$, then $x \in W^{cu}_{loc}$.*
5. *$W^u_{loc} = \{x \in B : g^n(x) \in B \text{ for all } n \leq 0 \text{ and } d(g^n(x), 0) \to 0 \text{ exponentially}\}$. Also, $g^{-1} : W^u_{loc} \to W^u_{loc}$ is a contraction map.*

If the gradient descent method remains close to a critical point for all time, then it is on the center-stable manifold. Note that $W^s_{\text{loc}} \subset W^{cs}_{\text{loc}}$.

Theorem 3.2. *Let f be a $C^2(\Omega)$ function where Ω is a forward invariant convex subset of $\mathbb{R}^d$ whose gradient has Lipschitz constant L, and let $\sigma(\cdot)$ denote the spectrum of a matrix. Consider the gradient descent method $g(x) = x - \alpha \nabla f(x)$ with $\alpha L \in (0, 2)$, and assume that the set*

$$\left\{ x \in \Omega \;\middle|\; \alpha^{-1} \in \sigma(\nabla^2 f(x)) \right\}$$

has measure zero and does not contain saddle points. Then the probability of gradient descent converging to a strict saddle, given one uniformly random initialization in Ω, is zero.

Proof. For simplicity of exposition, all sets are assumed to be in Ω; otherwise, one can either shrink the set or replace the set with the intersection with Ω (depending on the context).

First, we will show that g^{-1} maps null sets to null sets (in Ω), which follows from the assumption that g is C^1 and the set

$$\left\{x \in \Omega \;\middle|\; \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}$$

has measure zero. The map g is noninvertible only on the set

$$A := \left\{x \in \Omega \;\middle|\; \det(Dg(x)) = 0\right\},$$

which is equivalent to

$$\begin{aligned} A &= \left\{x \in \Omega \,\middle|\, 0 \in \sigma(Dg(x))\right\} \\ &= \left\{x \in \Omega \,\middle|\, 0 \in \sigma(I - \alpha\nabla^2 f(x))\right\} \\ &= \left\{x \in \Omega \,\middle|\, \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}. \end{aligned} \tag{3.1}$$

Note that if $\alpha L < 1$, then this set is measure zero by definition. For a point $x \in \Omega \setminus A$, we can find a neighborhood of x such that $\det(Dg(x)) \neq 0$ by continuity. By the inverse function theorem, g^{-1} is continuous differentiable. This implies that g maps sets of measure zero to sets of measure zero in $\Omega \setminus A$. To extend it to all of Ω, consider the following. Let $\{V_j\}_j$ be a collection of open neighborhoods that form a (countable) covering of $\Omega \setminus A$ such that $V_j \cap A = \emptyset$. Construct such a covering by first finding a neighborhood for each $x \in \Omega \setminus A$ that avoids A and then applying Lindelöf's lemma to find a countable subcovering. Given an arbitrary null set $U \subset \Omega$, we have

$$g^{-1}(U) \subset A \cup \left(\cup_j \; \left(V_j \cap g^{-1}(U)\right) \right).$$

The inverse function theorem can then be applied to each set $V_j \cap g^{-1}(U)$; therefore, since each set has measure zero, the countable union has zero measure. This implies that the set $g^{-1}(U)$ also has measure zero. Since U is arbitrary, this shows that g^{-1} sends null sets to null sets (within Ω).

Next, we want to show that all initializations that are mapped to degenerate points in A form a measure zero set. The set of all points in Ω which are iteratively mapped into A by g is equivalent to

$$\bigcup_{j=1}^{\infty} g^{-j}(A)$$

and has zero measure since it is the countable union of measure zero sets. By assumption, Ω is forward invariant; thus, initializations in Ω cannot lead to degenerate points outside of Ω. This implies that the probability of a random initialization in Ω mapping to a degenerate point is zero.

Finally, we want to show that the set of initializations that converge to a strict saddle point has zero measure. Let

$$x^0 \in \Omega \setminus \bigcup_{j=1}^{\infty} g^{-j}(A)$$

such that $\lim g^n(x^0)$ converges to a strict saddle x_k. Note that along this trajectory, $g^n(x^0)$ is not in A and thus is nondegenerate. Then, by the inverse function theorem and the assumption, it is a local C^1 diffeomorphism. Since g is continuously differentiable and nondegenerate at the strict saddle point x_k, there exists an open neighborhood $U(x_k)$ around x_k such that the spectrum of $Dg(x_k)$ is nonzero, and thus $A \cap U(x_k) = \emptyset$. For each strict saddle point, there exists a ball $B(x_k) \subset U(x_k)$ that satisfies the conditions in Theorem 3.1. The collection of such balls (over all strict saddle points)

$$\bigcup_k B(x_k)$$

are an open cover of the strict saddle points, so there exists a countable subcover, i.e.,

$$\bigcup_k x_k \in \bigcup_{\ell=1}^{\infty} B(x_\ell).$$

Thus, there exists an N such that

$$g^n(x^0) \in \bigcup_{\ell=1}^{\infty} B(x_\ell)$$

for all $n \geq N$. This implies that there exists an ℓ such that $g^n(x^0) \in B(x_\ell)$ for all $n \geq N$, and by Theorem 3.1, $g^n(x^0) \in W^{cs}_{loc}(x_\ell)$ for any $n \geq N$.

We will show that the set $W^{cs}_{loc}(x_\ell)$ has measure zero. By the strict saddle condition, we have that $Dg(x) = I - \alpha \nabla^2 f(x)$ has at least one eigenvalue with magnitude greater than 1; thus, the dimension of E^u is at least one, and therefore $\dim(W^{cs}_{\text{loc}}(x_\ell)) \leq d - 1$, and the Lebesgue measure of $W^{cs}_{\text{loc}}(x_\ell)$ is zero. Since $g^n(x^0) \in B(x_\ell)$ for any $n \geq N$, we have that

$$g^N(x^0) \in \bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell));$$

i.e., $g^N(x^0)$ is contained in the intersection of all domains which are mapped into the ball $B(x_\ell)$. The set

$$\bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell))$$

is contained in $W^{cs}_{loc}(x_\ell)$, so it has measure zero. Since

$$g^N(x^0) \in \bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell)),$$

we have that

$$x^0 \in g^{-N}\left(\bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell))\right).$$

The integer N depends on the initialization x^0 and the fixed point x_ℓ; thus, we must consider an arbitrary N. In particular, the backward map g^{-1} is in C^1; thus, the measure of

$$g^{-n}\left(\bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell))\right)$$

is zero for all $n \geq 0$. Note that a countable union of measure zero sets are measure zero, so the set

$$\mathcal{S} = \bigcup_{\ell=0}^{\infty}\bigcup_{n=0}^{\infty} g^{-n}\left(\bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell))\right)$$

has measure zero as well. The set $\mathcal{S}$ contains all points in

$$\Omega \setminus \bigcup_{j=1}^{\infty} g^{-j}(A)$$

which converge to strict saddles; thus, the measure of all points in Ω that converge to a strict saddle is zero. ∎

As was shown in the proof, the condition that the set $\left\{x \in \Omega \;\middle|\; \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}$ has measure zero implies that g^{-1} has the Luzin N property over sets in Ω. The following is a direct result of Theorem 3.2 for the step-size $\alpha L = 1$.

Corollary 3.3. *Let f be a $C^2(\Omega)$ function where Ω is a forward invariant convex subset of $\mathbb{R}^d$ whose gradient has Lipschitz constant L. Consider the gradient descent method $g(x) = x - L^{-1}\nabla f(x)$, and assume that the set where $\sigma(\nabla^2 f(x))$ achieves its maximum has measure zero and does not contain saddles. Then the probability of gradient descent converging to a strict saddle, given one uniformly random initialization in Ω, is zero.*

Example 2.3 shows that the measure zero assumption on the degenerate set is necessary. In addition, note that the results above do not assume that the strict saddles are isolated.

3.1. Weaker condition: Positive Lipschitz restriction.

Define

$$\ell(x) := \max_{1\leq j\leq d} \max(\lambda_j(x), 0),$$

where λ_j is an eigenvalue of the Hessian, and let L_+ be the Lipschitz constant of the positive part:

$$L_+ = \sup_{x\in\Omega} \ell(x).$$

Then we can show that control of L_+ is sufficient for avoiding strict saddles, although it may not imply convergence to minima.

Example 3.4. Consider the objective function $f(x,y) := Q(x) + \frac{1}{b}y^2$, where Q is defined as the even function with

$$Q(x) = \begin{cases} a\cos(x) & \text{if } \ x \le \tilde{x}, \\ \frac{1}{b}\left(x - \tilde{x} - \frac{ab}{2}\sin(\tilde{x})\right)^2 - \frac{2}{b} - \frac{a^2 b}{4}\sin^2(\tilde{x}) & \text{if } \ x > \tilde{x} \end{cases}$$

and where $\tilde{x} = \arccos(-\frac{2}{ab})$ with $ab \ge 2$ and a and b positive (thus, $\tilde{x} \in [\pi/2, \pi]$). The function has three critical points: $(0,0)$ a strict saddle and two minima defined at $\pm(\tilde{x} + \frac{ab}{2}\sin(\tilde{x}), 0)$. The Hessian is diagonal with eigenvalues given by $Q''(x)$ and $\frac{2}{b}$. A Lipschitz constant is $L = a$ and is obtained at $x = 0$, and the positive Lipschitz constant is $L_+ = \frac{2}{b}$.

Consider the gradient descent method with $\alpha = L_+^{-1} = \frac{b}{2}$; then $y^n = 0$ for all $n > 1$. The iterative map for x^n is defined by

$$x^{n+1} = \begin{cases} x^n + \frac{ab}{2}\sin(x^n) & \text{if } \ |x| \le \tilde{x}, \\ \tilde{x} + \frac{ab}{2}\sin(\tilde{x}) & \text{if } \ x > \tilde{x}, \\ -\tilde{x} - \frac{ab}{2}\sin(\tilde{x}) & \text{if } \ x < -\tilde{x}. \end{cases}$$

For points in $0 < |x| < \tilde{x}$, the map expands away from zero (since in $|x| < \pi, \sin(x)$ and x share the same sign). Therefore, points in $0 < |x| < \tilde{x}$ will flow to $|x| \ge \tilde{x}$. For any point $|x| \ge \tilde{x}$, the map will converge (in one step) to $\pm(\tilde{x} + \frac{ab}{2}\sin(\tilde{x}))$. This shows that even if L/L_+ is arbitrary large, control of L_+ will be sufficient to avoid the strict saddle point.

To visualize this, in Figure 3.1, we set $\alpha = L_+^{-1}$ and plot the gradient descent trajectories, sampled from a set of random initial states. As expected, the gradient descent path moves away from the stable manifold of the saddle (i.e., $x = 0$) and will not converge to the strict saddle.

Recall that $Dg(x) = I - \alpha\nabla^2 f(x)$, and if we assume $\alpha L_+ < 1$, then all eigenvalues of $Dg(x)$ are strictly positive. Since the spectrum of $Dg(x)$ is strictly positive and $g \in C^1$, by the inverse function theorem, g is a diffeomorphism under the positive Lipschitz condition. Proposition 3.5 and Corollary 3.6 are extensions of theorems from [14, 21] using the techniques from this section and from Theorem 3.2. In particular, we have the following refinement.

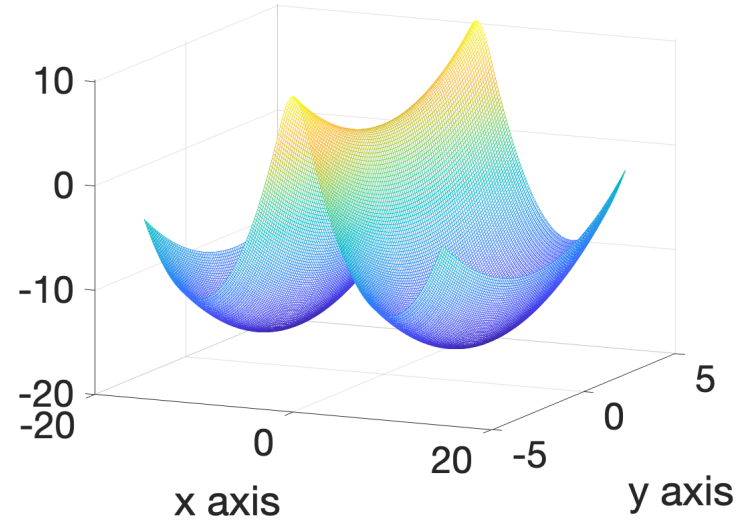
Proposition 3.5. *Let $f \in C^2(\Omega)$, where Ω is a forward invariant convex subset of $\mathbb{R}^d$ whose gradient has positive Lipschitz constant L_+. Consider the gradient descent method $g(x) = x - \alpha\nabla f(x)$ with $\alpha L_+ \in (0,1)$. Then the probability of gradient descent converging to a strict saddle, given one uniformly random initialization in Ω, is zero.*

To extend this result beyond $\alpha L_+ < 1$, we add the assumption from Theorem 3.2.

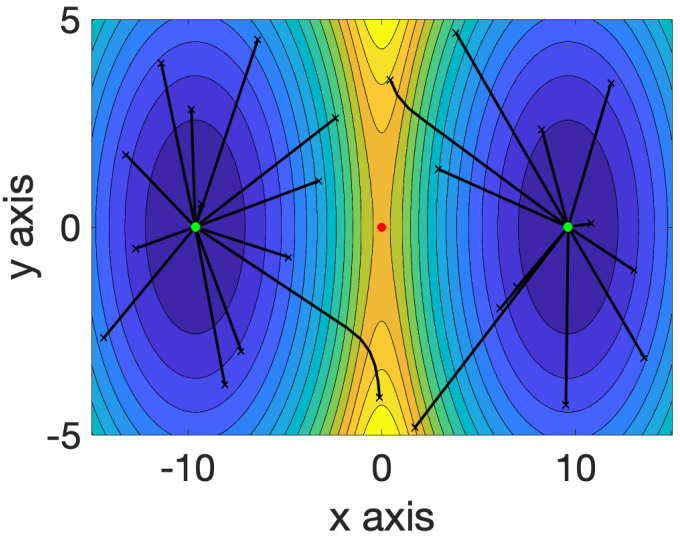
Corollary 3.6. *Let $f \in C^2(\Omega)$, where Ω is a forward invariant convex subset of $\mathbb{R}^d$ whose gradient has Lipschitz constant L_+. Consider the gradient descent method $g(x) = x - \alpha\nabla f(x)$ with $\alpha L_+ \in (0,2)$, and assume that the set $\left\{x \in \Omega \mid \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}$ has measure zero and does not contain saddles. Then the probability of gradient descent converging to a strict saddle, given one uniformly random initialization in Ω, is zero.*

3.2. Varying steps-size. In some applications, the step-size of gradient descent changes between iterations. We consider a variable step-size gradient descent method,

$$x^{n+1} = x^n - \alpha^n \nabla f(x^n),$$



(a) Surface Plot



(b) Random Gradient Descent Trajectories

Figure 3.1. *The plots correspond to the surface* (a) *of the objective function from Example* 3.4 *and the associated gradient descent trajectories* (b) *from a set of randomly sampled initial conditions. The red dot is the strict saddle point, and the green dots are the local minimizers. Only trajectories whose initial x value is zero will converge to the saddle point. Note that points near the line $x = 0$ but not on it will move away from the stable manifold of the strict saddle.*

where $\alpha^n > 0$. By augmenting the iterative system with the step-size as an additional variable, we can apply the results of Theorem 3.2 to show that the iterations avoid strict saddles.

Corollary 3.7. *Let f be a $C^2(\Omega)$ function where Ω is a forward invariant convex subset of $\mathbb{R}^d$ whose gradient has Lipschitz constant L. Consider the gradient descent method with varying*

step-size satisfying that $\alpha^{n+1} = h(\alpha^n)$*, where* $h \in C^1$ *is a strictly decreasing contractive map over the interval* $\mathcal{I}$ *containing the unique fixed point* $\alpha^* > 0$ *and the initial step-size* α^0*. If* $\alpha^0 L \in (0, 2)$*,* $\alpha^0 \geq \alpha^*$*, and the set*

$$\bigcup_{L^{-1} \leq \alpha \leq \alpha^0} \left\{ x \in \Omega \mid \alpha^{-1} \in \sigma(\nabla^2 f(x)) \right\}$$

has measure zero and does not contain saddle points, then the probability of gradient descent converging to a strict saddle, given one uniformly random initialization in Ω*, is zero.*

Proof. By augmenting the iterations with the time-step variable, the gradient descent method becomes

$$\begin{cases} x^{n+1} = x^n - \alpha^n \nabla f(x^n), \\ \alpha^{n+1} = h(\alpha^n) \end{cases}$$

and can be analyzed via Theorem 3.2. The updated function $g(x, \alpha)$ is defined by $g(x, \alpha) = [x - \alpha \nabla f(x),\ h(\alpha)]^T$, and its Jacobian is given by

$$Dg(x, \alpha) = \begin{bmatrix} I - \alpha \nabla^2 f(x) & -\nabla f(x) \\ 0_{1 \times n} & h'(\alpha) \end{bmatrix}.$$

Since the Jacobian is "block-upper-triangular," its eigenvalues are the eigenvalues $I - \alpha \nabla^2 f(x)$ and $h'(\alpha)$. Since h' is negative, the degeneracy in g must come from x. In addition, by the assumptions on h, α^n converges to α^* for any initialization of $\alpha^0 \geq \alpha^*$.

Define the set $\Omega_1 = \Omega \times \mathcal{I}$, and let $A \subset \Omega_1$ denote the set of points where g is noninvertible, i.e.,

$$\begin{aligned} A &= \left\{ (x, \alpha) \in \Omega_1 \,\middle|\, 0 \in \sigma(Dg(x)) \right\} \\ &= \left\{ x \in \Omega,\ \alpha \in \mathcal{I} \,\middle|\, 0 \in \left\{ \sigma(I - \alpha \nabla^2 f(x)), h'(\alpha) \right\} \right\} \qquad (3.2) \\ &= \left\{ x \in \Omega, \alpha \in \mathcal{I} \,\middle|\, \alpha^{-1} \in \sigma(\nabla^2 f(x)) \right\} \qquad (3.3) \\ &= \bigcup_{L^{-1} \leq \alpha \leq \alpha^0} \left\{ x \in \Omega \mid \alpha^{-1} \in \sigma(\nabla^2 f(x)) \right\}. \end{aligned}$$

By assumption, A has measure zero.

The set Ω_1 is a convex subset of $\mathbb{R}^{d+1}$. By assumption, the function $g_1(x, \alpha) = x - \alpha \nabla f(x)$ is forward invariant on Ω_1. In addition, $g_2(x, \alpha) = h(\alpha)$ is a contractive map ($|h'(\alpha)| < 1$); thus, $h(\mathcal{I}_1) \subset \mathcal{I}$. Therefore, g is forward invariant on Ω_1.

Let

$$(x^0, \alpha^0) \in \Omega_1 \setminus \bigcup_{j=1}^{\infty} g^{-j}(A)$$

such that $\lim g^n(x^0, \alpha^0)$ converges to a strict saddle (x, α^*) (the fixed point for α is unique). The map g is continuously differentiable and nondegenerate at (x, α^*); thus, there exists an open neighborhood around (x, α^*) characterized by the product space of an open neighborhood $U(x)$ around x and an open interval $S(\alpha^*)$ (which holds by the odd extension of h), where the spectrum of $Dg(x)$ is nonzero; thus, $A \cap U(x) = \emptyset$. The rest follows from Theorem 3.2. ∎

The theorem above holds (trivially) if $\alpha^0 < L$. If the set of step-sizes is discrete, we can simplify the results.

Corollary 3.8. *Let f be a $C^2(\Omega)$ function where Ω is a forward invariant convex subset of $\mathbb{R}^d$ whose gradient has Lipschitz constant L. Consider the gradient descent method with a finite staircase of decreasing step-sizes; i.e., α^n is a piecewise constant function of n with finitely many jumps. If $\alpha^n L \in (0,2)$ for all n and the set $\left\{x \in \Omega \,\middle|\, \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}$ has measure zero for each α^n and does not contain saddle points, then the probability of gradient descent converging to a strict saddle, given one uniformly random initialization in Ω, is zero.*

Proof. Consider the case $\alpha^n = \alpha_1$ for $n \le N_1$ and $\alpha^n = \alpha_2$ for $n > N_1$. Let g_i be the gradient descent method with step-size α_i, $i = 1, 2$.

The maps g_i are C^1 and are noninvertible only on the set $A_i := \left\{x \in \Omega \,\middle|\, \det(Dg_i(x)) = 0\right\}$ (respectively), which is equivalent to

$$A_i = \left\{x \in \Omega \,\middle|\, \alpha_i^{-1} \in \sigma(\nabla^2 f(x))\right\}.$$

Following the proof of Theorem 3.2, g_i^{-1} maps null sets to null sets (within Ω). Consider the set $A = \cup_i A_i$, which is a null set since it is a finite union of null sets. All points in Ω that are mapped to A by g_i (for any i) is equivalent to the set

$$Q = \bigcup_i \bigcup_{j=1}^{\infty} g_i^{-j}(A).$$

Each A_i is a null set, so $g_i^{-j}(A)$ are null sets. The set Q is a countable union of null sets; thus, Q has measure zero.

Let $x^0 \in \Omega \setminus Q$ such that the two-step staircase gradient descent method converges to a strict saddle x. This can occur by two distinct scenarios: (i) $g_1^n(x^0)$ converges to x within N_1 steps, or (ii) $g_2^{n-N_1}(g_1^N(x^0))$ converges to x with $n > N_1$. For case (i), using the proof of Theorem 3.2, the set of points in $\Omega \setminus Q$ which converge to a strict saddle under g_1 is measure zero.

For case (ii), by assumption, $x_0 \notin Q$, so $x^{N_1} := g_1^{N_1}(x^0) \notin Q$; i.e., along the trajectory $g_2^{n-N_1}(x^{N_1})$ for $n > N_1$, g_2 is nondegenerate and a local C^1 diffeomorphism.

As before, we can show that there exists a (sufficiently large) N such that

$$g_2^n(x^{N_1}) = \bigcup_{\ell=1}^{\infty} B(x_\ell)$$

for all $n \ge N$, and thus there is an ℓ such that $g_2^n(x^{N_1}) \in B(x_\ell)$ for all $n \ge N$ and $g_2^n(x^{N_1}) \in W^{cs}_{loc}(x_\ell)$ for any $n \ge N$. This also implies that

$$g_2^N(x^{N_1}) \in \bigcap_{j=0}^{\infty} g_2^{-j}(B(x_\ell)),$$

which is measure zero since it is contained in $W^{cs}_{loc}(x_\ell)$. Since $g_2^N(x^{N_1}) \in \bigcap_{j=0}^{\infty} g^{-j}(B(x_\ell))$, we can show that

$$x^{N_1} \in g_2^{-N}\left(\bigcap_{j=0}^{\infty} g_2^{-j}(B(x_\ell))\right),$$

$$x^0 \in g_1^{-N_1}\left(g_2^{-N}\left(\bigcap_{j=0}^{\infty} g_2^{-j}(B(x_\ell))\right)\right).$$

The set

$$\mathcal{S} = \bigcup_{\ell=0}^{\infty}\bigcup_{n=0}^{\infty} g_1^{-N_1}\left(g_2^{-n}\left(\bigcap_{j=0}^{\infty} g_2^{-j}(B(x_\ell))\right)\right)$$

contains all points in $\Omega \setminus Q$ which converge to strict saddles after N_1 iterations. The set $\mathcal{S}$ has zero measure since g_i^{-1} map null sets to null sets and $\mathcal{S}$ is the countable union of null sets. Therefore, the probability of case (ii) occurring is zero.

This can be generalized to finitely many discrete step-sizes since the arguments related to the invertibility of all g_i continue to hold for countable unions of null sets. ∎

4. Robustness of assumptions. If we consider the space of real analytic functions $C^\omega(\Omega)$ equipped with the sup-norm, then we can show that a generic function will satisfy the conditions in the previous results. Since we consider a generic function $f \in C^\omega(\Omega)$, the Lipschitz constant should directly depend on the function, which is denoted by L_f. Let μ denote the Lebesgue measure.

Theorem 4.1. *Let Ω be a (bounded) forward invariant convex subset of $\mathbb{R}^d$. Let L_f denote the Lipschitz constant of the gradient of f. For a given $c \in (0,2)$, the set of functions in $C^\omega(\Omega)$ with*

$$\mu\left(\left\{x \in \Omega \;\middle|\; cL_f^{-1} \in \sigma(\nabla^2 f(x))\right\}\right) > 0$$

is meager.

Proof. If $c \in (0,1)$, the condition is satisfied, so we only need to consider $c \in [1,2)$. Let $\alpha = cL_f^{-1}$, and define the set U by

$$U := \left\{f \in C^\omega(\Omega) \;\middle|\; \mu\left(\left\{x \in \Omega \;\middle|\; \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}\right) > 0\right\}.$$

To find an equivalent formulation of this set, define $S := \left\{x \in \Omega \,\middle|\, \alpha^{-1} \in \sigma(\nabla^2 f(x))\right\}$, and note that the function $g : \mathbb{R}^d \to \mathbb{R}$, defined by $g(x) := \det(\nabla^2 f(x) - \alpha^{-1} I)$, is zero on S. Since g is analytic and zero on a set of positive measure, $g(x) = 0$ on all of Ω [16]. Therefore,

$$\begin{aligned} U &= \left\{f \in C^\omega(\Omega) \;\middle|\; \det(\nabla^2 f(x) - \alpha^{-1} I) = 0 \text{ on some measure positive set}\right\} \\ &= \left\{f \in C^\omega(\Omega) \;\middle|\; \det(\nabla^2 f(x) - \alpha^{-1} I) = 0 \;\;\forall\;\; x \in \Omega\right\}. \end{aligned}$$

To show that U is meager, we will show that U is closed and has empty interior. The complement set (based on analyticity) can be written as

$$U^c = \left\{f \in C^\omega(\Omega) \;\middle|\; \det(\nabla^2 f(x) - \alpha^{-1} I) \neq 0 \text{ a.e. in } \Omega\right\}.$$

The function $G : C^\omega(\Omega) \to [0, \infty)$, defined by

$$G(f) = \|\det(\nabla^2 f(x) - \alpha^{-1} I)\|_{L^1(\Omega,\mu)} = \int_\Omega \left|\det(\nabla^2 f(x) - \alpha^{-1} I)\right| \, d\mu(x),$$

is positive on U^c; i.e., $G(f) > 0$ for all $f \in U^c$ since the integrand is nonzero almost everywhere. For all $f \in U$, the function is zero, $G(f) = 0$; therefore, U^c can be characterized as

$$U^c = \left\{ f \in C^\omega(\Omega) \;\middle|\; G(f) \in (0, \infty) \right\} = G^{-1}\left((0, \infty)\right);$$

i.e., U^c is the preimage of the open set $(0, \infty)$ under the continuous function G, so U^c is open.

Next, we will show that U has empty interior. If $f \in U$, then $\det(\nabla^2 f(x) - \alpha^{-1} I) = 0$ on Ω. By picking any point $x_0 \in \Omega$, we will perturb the eigenspace globally as follows. The spectrum of the Hessian contains α^{-1}, i.e., $\alpha^{-1} \in \sigma(\nabla^2 f(x_0))$; thus there exists a collection of m-orthogonal eigenvectors, denoted by the matrix $V \in \mathbb{R}^{d \times m}$, that correspond to the eigenvalue α^{-1}. The perturbed function is defined as

$$\tilde{f}(x) := f(x) + \frac{\epsilon}{2}(x - x^0)^T (V V^T)(x - x^0),$$

and thus $\nabla^2 \tilde{f}(x) = \nabla^2 f(x) + \epsilon V V^T$. The function $\tilde{f}$ perturbs the eigenspace of f at x^0 in the direction of the eigenspace corresponding to the eigenvalue α^{-1}. Thus, we have that $\alpha^{-1} \notin \sigma(\nabla^2 \tilde{f}(x_0))$ for small enough $\epsilon > 0$. Since the Hessian is Hermitian, the (ordered) eigenvalues are Lipschitz continuous with respect to the Hessian (by Weyl's inequality). The eigenvalues are sufficiently continuous to show that there is a neighborhood $S_\delta = \{\|x - x_0\| \le \delta\}$ with positive measure δ (sufficiently small) such that $\det(\nabla^2 \tilde{f}(x) - \alpha^{-1} I) \neq 0$ on S_δ, and thus the measure of the set such that $\det(\nabla^2 \tilde{f}(x) - \alpha^{-1} I) = 0$ must be zero. Therefore, the perturbed function is not in U, i.e., $\tilde{f} \in C^\omega(\Omega) \setminus U$; however, it can be made arbitrary close:

$$||f - \tilde{f}||_{\sup} \le C(\operatorname{diam}(\Omega)) \, \epsilon.$$

Hence, U has empty interior. ∎

This shows that, in some sense, the assumptions from Theorem 3.2 and the related results are generic and indeed describe a large class of objective functions.

5. Discussion. We present several theoretical results on the conditions which guarantee that the gradient descent method will avoid a strict saddle. The results utilize the center manifold theorem to establish the size of the attracting sets and measure theoretic arguments to show that the iterative maps satisfy the Luzin N condition. Our results answer an open question about the step-size posed in [14, 13], namely, that previous claims hold for $\alpha < 2L^{-1}$ with the additional assumption that the iterative map does not degenerate on nonnull sets. We show that without the additional assumption, one can construct counterexamples. These results also hold for the gradient descent method with (fixed) learning rate schedules.

Extensions and applications. The theoretical results here extend readily to other first-order methods, for example, the proximal gradient descent, block coordinate descent, etc. [13]. Although the results are for uniformly random initial data, they can be easily extended to

other random sampling measures. Additionally, using the Łojasiewicz gradient inequality [1], one may be able to prove that if the set of critical points only contains local minima and strict saddles, then the gradient descent method converges to local minima with the extended step-sizes [14].

Limitations. This paper does not directly address the convergence of gradient descent to global minimizers or the behavior near local minimizers. In particular, the step-size bounds presented here may be too large for convergence when applied to a particular model. Additionally, it was shown in [7] that the gradient descent method can take exponential time to escape a saddle, but the likelihood or predictability of such phenomena for a particular model or application is an open question. Finally, our results on varying step-sizes utilized a fixed learning rate schedule. A line search or adaptive time-stepping method may be able to avoid saddles with weaker restrictions on α.

6. Acknowledgment. The authors would like to thank Stephen Wright for his helpful comments.

REFERENCES

[1] P.-A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM J. Optim., 16 (2005), pp. 531–547.
[2] A. Anandkumar and R. Ge, *Efficient approaches for escaping higher order saddle points in non-convex optimization*, Proc. Mach. Learn. Res., 49 (2016), pp. 81–102.
[3] S. Bhojanapalli, B. Neyshabur, and N. Srebro, *Global optimality of local search for low rank matrix recovery*, in Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press, 2016, pp. 3873–3881.
[4] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, *Accelerated methods for nonconvex optimization*, SIAM J. Optim., 28 (2018), pp. 1751–1772.
[5] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust Region Methods*, Vol. 1, SIAM, Philadelphia, 2000.
[6] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, in Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press, 2014, pp. 2933–2941.
[7] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, *Gradient descent can take exponential time to escape saddle points*, in Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press, 2017, pp. 1067–1077.
[8] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points—Online stochastic gradient for tensor decomposition*, Proc. Mach. Learn. Res., 40 (2015), pp. 797–842.
[9] R. Ge, C. Jin, and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, Proc. Mach. Learn. Res., 70 (2017), pp. 1233–1242.
[10] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, in Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press, 2016, pp. 2973–2981.
[11] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, Proc. Mach. Learn. Res., 70 (2017), pp. 1724–1732.
[12] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, *On large-batch training for deep learning: Generalization gap and sharp minima*, in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 2019.
[13] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid strict saddle points*, Math. Program., 176 (2019), pp. 311–337.
[14] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, Proc. Mach. Learn. Res., 49 (2016), pp. 1246–1257.
[15] K. Y. Levy, *The Power of Normalization: Faster Evasion of Saddle Points*, preprint, https://arxiv.org/abs/1611.04831, 2016.

[16] B. S. Mityagin, *The zero set of a real analytic function*, Mat. Zametki, 107 (2020), pp. 473–475.
[17] J. J. Moré and D. C. Sorensen, *On the use of directions of negative curvature in a modified Newton method*, Math. Program., 16 (1979), pp. 1–20.
[18] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Vol. 87, Springer, New York, 2013.
[19] Y. Nesterov and B. T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
[20] M. O'Neill and S. J. Wright, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, Math. Program., (2017), pp. 1–25.
[21] I. Panageas and G. Piliouras, *Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions*, in 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Wadern, Germany, 2017.
[22] R. Pascanu, Y. N. Dauphin, S. Ganguli, and Y. Bengio, *On the Saddle Point Problem for Non-Convex Optimization*, preprint, https://arxiv.org/abs/1405.4604, 2014.
[23] R. Pemantle, *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18 (1990), pp. 698–712.
[24] S. Reddi, M. Zaheer, S. Sra, B. Poczos, F. Bach, R. Salakhutdinov, and A. Smola, *A generic approach for escaping saddle points*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, A. Storkey and F. Perez-Cruz, eds., Proceedings of Machine Learning Research 84, Playa Blanca, Lanzarote, Canary Islands, April 9–11, 2018, PMLR, pp. 1233–1242, http://proceedings.mlr.press/v84/reddi18a.html.
[25] C. W. Royer and S. J. Wright, *Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization*, SIAM J. Optim., 28 (2018), pp. 1448–1477.
[26] M. Shub, *Global Stability of Dynamical Systems*, Springer, New York, 2013.
[27] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, *Theoretical insights into the optimization landscape of over-parameterized shallow neural networks*, IEEE Trans. Inform. Theory, 65 (2019), pp. 742–769.
[28] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere* I: *Overview and the geometric picture*, IEEE Trans. Inform. Theory, 63 (2016), pp. 853–884.
[29] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere* II: *Recovery by Riemannian trust-region method*, IEEE Trans. Inform. Theory, 63 (2017), pp. 885–914.
[30] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Found. Comput. Math., 18 (2018), pp. 1131–1198.