JMLR: Workshop and Conference Proceedings vol 49:1–12, 2016

Gradient Descent Only Converges to Minimizers

Jason D. Lee LEE715@MARSHALL.USC.EDU
Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089

Max Simchowitz MSIMCHOW@BERKELEY.EDU

Michael I. Jordan JORDAN@CS.BERKELEY.EDU

Benjamin Recht RECHT@BERKELEY.EDU
Department of Electrical Engineering and Computer Science
UC Berkeley
Berkeley, CA 94720

Abstract

We show that gradient descent converges to a local minimizer, almost surely with random initialization. This is proved by applying the Stable Manifold Theorem from dynamical systems theory.
Keywords: Gradient descent, saddle points, local minimum, non-convex

1. Introduction

Saddle points have long been regarded as a tremendous obstacle for continuous optimization. There are many well known examples when worst case initialization of gradient descent provably converge to saddle points (Nesterov, 2004, Section 1.2.3), and hardness results which show that finding even a *local* minimizer of non-convex functions is NP-Hard in the worst case (Murty and Kabadi, 1987). However, such worst-case analyses have not daunted practitioners, and high quality solutions of continuous optimization problems are readily found by a variety of simple algorithms. Building on tools from the theory of dynamical systems, this paper demonstrates that, under very mild regularity conditions, saddle points are indeed of little concern for the gradient method.

More precisely, let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable, and consider the classic gradient method with constant step size α:

$$x_{k+1} = x_k - \alpha \nabla f(x_k). \tag{1}$$

We call x a critical point of f if $\nabla f(x) = 0$, and say that f satisfies the *strict saddle property* if each critical point 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. Informally, we prove:

> If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and satisfies the strict saddle property, then gradient descent (Equation 1) with a random initialization and sufficiently small constant step size almost surely converges to a local minimizer.

Here, by sufficiently small, we simply mean less than the inverse of the Lipschitz constant of the gradient. As we discuss below, such step sizes are standard for the gradient method. We remark that the strict saddle assumption is necessary in the worst case, due to hardness results regarding testing the local optimality of functions whose Hessians are highly degenerate at critical points (e.g, quartic polynomials) (Murty and Kabadi, 1987).

1.1. Related work

Prior work has show that first-order descent methods can circumvent strict saddle points, provided that they are augmented with unbiased noise whose variance is sufficiently large along each direction. For example, Pemantle (1990) establishes convergence of the Robbins-Monro stochastic approximation to local minimizers for strict saddle functions. More recently, Ge et al. (2015) give quantitative rates on the convergence of noise-added stochastic gradient descent to local minimizers, for strict saddle functions. The condition that the noise have large variance along all directions is often not satisfied by the randomness which arises in sample-wise or coordinate-wise stochastic updates. In fact, it generally requires that additional, near-isotropic noise be added at each iteration, which yields convergence rates that depend heavily on problem parameters like dimension. In contrast, our results hold for the simplest implementation of gradient descent and thus do not suffer from the slow convergence associated with adding high-variance noise to each iterate.

But is this strict saddle property reasonable? Many works have answered in the affirmative by demonstrating that many objectives of interest do in fact satisfy the "strict saddle" property: PCA, a fourth-order tensor factorization (Ge et al., 2015), formulations of dictionary learning (Sun et al., 2015b,a) and phase retrieval (Sun et al., 2016).

To obtain provable guarantees, the authors of Sun et al. (2015b,a) and Sun et al. (2016) adopt trust-region methods which leverage Hessian information in order to circumvent saddle points. This approach joins a long line of related strategies, including: a modified Newton's method with curvilinear line search (Moré and Sorensen, 1979), the modified Cholesky method (Gill and Murray, 1974), trust-region methods (Conn et al., 2000), and the related cubic regularized Newton's method (Nesterov and Polyak, 2006), to name a few. Specialized to deep learning applications, Dauphin et al. (2014); Pascanu et al. (2014) have introduced a saddle-free Newton method.

Unfortunately, such curvature-based optimization algorithms have a per-iteration computational complexity which scales quadratically or even cubically in the dimension d, rendering them unsuitable for optimization of high dimensional functions. In contrast, the complexity of an iteration of gradient descent is linear in dimension. We also remark that the authors of Sun et al. (2016) empirically observe gradient descent with 100 random initializations on the phase retrieval problem reliably converges to a local minimizer, and one whose quality matches that of the solution found using more costly trust-region techniques.

More broadly, many recent works have shown that gradient descent plus smart initialization provably converges to the global minimum for a variety of non-convex problems: such settings include matrix factorization (Keshavan et al., 2009; Zhao et al.) , phase retrieval (Candes et al., 2015; Cai et al., 2015), dictionary learning (Arora et al., 2015), and latent-variable models (Zhang et al., 2014; Belkin et al., 2014). While our results only guarantee convergence to local minimizers, they eschew the need for complex and often computationally prohibitive initialization procedures.

Finally, some preliminary results have shown that there are settings in which if an algorithm converges to a saddle point it necessarily has a small objective value. For example, Choromanska

et al. (2014) studies the loss surface of a particular Gaussian random field as a proxy for understanding the objective landscape of deep neural nets. The results leverage the Kac-Rice Theorem (Adler and Taylor, 2009; Auffinger et al., 2013), and establish that that critical points with more positive eigenvalues have lower expected function value, often close to that of the global minimizer. We remark that functions drawn from this Gaussian random field model share the strict saddle property defined above, and so our results apply in this setting. On the other hand, our results are considerably more general, as they do not place stringent generative assumptions on the objective function f.

1.2. Organization

The rest of the paper is organized as follows. Section 2 introduces the notation and definitions used throughout the paper. Section 3 provides an intuitive explanation for why it is unlikely that gradient descent converges to a saddle point, by studying a non-convex quadratic and emphasizing the analogy with power iteration. Section 4 states our main results which guarantee gradient descent converges to only local minimizers, and also establish rates of convergence depending on the local geometry of the minimizer. The primary tool we use is the local stable manifold theorem, accompanied by inversion of gradient descent via the proximal point algorithm. Finally, we conclude in Section 5 by suggesting several directions of future work.

2. Preliminaries

Throughout the paper, we will use f to denote a real-valued function in C^2, the space of twice-continuously differentiable functions, and g to denote the corresponding gradient map with step size α,

$$g(x) = x - \alpha \nabla f(x). \tag{2}$$

The Jacobian of g is given by $Dg(x)_{ij} = \frac{\partial g_i}{\partial x_j}(x)$, or $Dg(x) = I - \alpha \nabla^2 f(x)$. In addition to being C^2, our main regularity assumption on f is that it has a Lipschitz gradient:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 .$$

The k-fold composition of the gradient map $g^k(x)$ corresponds to performing k steps of gradient descent initialized at x. The iterates of gradient descent will be denoted $x_k := g^k(x_0)$. All the probability statements are with respect to ν, the distribution of x_0, which we assume is absolutely continuous with respect to Lebesgue measure.

A fixed point of the gradient map g is a critical point of the function f. Critical points can be saddle points, local minima, or local maxima. In this paper, we will study the critical points of f via the fixed points of g, and then apply dynamical systems theory to g.

Definition 1

1. *A point x^* is a critical point of f if it is a fixed point of the gradient map $g(x^*) = x^*$, or equivalently $\nabla f(x^*) = 0$.*
2. *A critical point x^* is isolated if there is a neighborhood U around x^*, and x^* is the only critical point in U.*

3. *A critical point is a local minimum if there is a neighborhood U around x^* such that $f(x^*) \leq f(x)$ for all $x \in U$, and a local maximum if $f(x^*) \geq f(x)$.*

4. *A critical point is a saddle point if for all neighborhoods U around x^*, there are $x, y \in U$ such that $f(x) \leq f(x^*) \leq f(y)$.*

As mentioned in the introduction, we will be focused on saddle points that have directions of strictly negative curvature. This notion is made precise by the following definition.

Definition 2 (Strict Saddle) *A critical point x^* of f is a strict saddle if $\lambda_{\min}(\nabla^2 f(x^*)) < 0$.*

Since we are interested in the attraction region of a critical point, we define the stable set.

Definition 3 (Global Stable Set) *The global stable set $W^s(x^*)$ of a critical point x^* is the set of initial conditions of gradient descent that converge to x^*:*

$$W^s(x^*) = \{x : \lim_k g^k(x) = x^*\}.$$

3. Intuition

To illustrate why gradient descent does not converge to saddle points, consider the case of a non-convex quadratic, $f(x) = \frac{1}{2}x^T H x$. Without loss of generality, assume $H = \mathbf{diag}(\lambda_1, ..., \lambda_n)$ with $\lambda_1, ..., \lambda_k > 0$ and $\lambda_{k+1}, \ldots, \lambda_n < 0$. $x^* = 0$ is the unique critical point of this function and the Hessian at x^* is H. Note that gradient descent initialized from x_0 has iterates

$$x_{k+1} = \sum_{i=1}^{n} (1 - \alpha\lambda_i)^{k+1} \langle e_i, x_0 \rangle e_i \,.$$

where e_i denote the standard basis vectors. This iteration resembles power iteration with the matrix $I - \alpha H$.

The gradient method is guaranteed to converge with a constant step size provided $0 < \alpha < \frac{2}{L}$ (Nesterov, 2004). For this quadratic f, L is equal to $\max |\lambda_i|$. Suppose $\alpha < 1/L$, a slightly stronger condition. Then we will have $(1 - \alpha\lambda_i) < 1$ for $i \leq k$ and $(1 - \alpha\lambda_i) > 1$ for $i > k$. If $x_0 \in E_s := \mathrm{span}(e_1, \ldots, e_k)$, then x_k converges to the saddle point at 0 since $(1 - \alpha\lambda_i)^{k+1} \to 0$. However, if x_0 has a component outside E_s then gradient descent diverges to ∞. For this simple quadratic function, we see that the global stable set (attractive set) of 0 is the subspace E_s. Now, if we choose our initial point at random, the probability of that point landing in E_s is zero.

As an example of this phenomena for a non-quadratic function, consider the following example from (Nesterov, 2004, Section 1.2.3). Letting $f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$, the corresponding gradient mapping is

$$g(x) = \begin{bmatrix} (1 - \alpha)x \\ (1 + \alpha)y - \alpha y^3 \end{bmatrix}.$$

The critical points are

$$z_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad z_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad z_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The points z_2 and z_3 are isolated local minima, and z_1 is a saddle point.

Gradient descent initialized from any point of the form $\begin{bmatrix} x \\ 0 \end{bmatrix}$ converges to the saddle point z_1. Any other initial point either diverges, or converges to a local minimum, so the stable set of z_1 is the x-axis, which is a zero measure set in $\mathbf{R}^2$. By computing the Hessian,

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

we find that $\nabla^2 f(z_1)$ has one positive eigenvalue with eigenvector that spans the x-axis, thus agreeing with our above characterization of the stable set. If the initial point is chosen randomly, there is zero probability of initializing on the x-axis and thus zero probability of converging to the saddle point z_1.

In the general case, the local stable set $W^s_{loc}(x^*)$ of a critical point x^* is well-approximated by the span of the eigenvectors corresponding to positive eigenvalues. By an application of Taylor's theorem, one can see that if the initial point x_0 is uniformly random in a small neighborhood around x^*, then the probability of initializing in the span of these eigenvectors is zero whenever there is a negative eigenvalue. Thus, gradient descent initialized at x_0 will leave the neighborhood. The primary difficulty is that x_0 is randomly distributed over the entire domain, not a small neighborhood around x^*, and Taylor's theorem does not provide any global guarantees.

However, the global stable set can be found by inverting the gradient map via g^{-1}. Indeed, the global stable set is precisely $\cup_{k=0}^{\infty} g^{-k}(W^s_{loc}(x^*))$. This follows because if a point x converges to x^*, then for some sufficiently large k it must enter the local stable set. That is, x converges to x^* if and only if $g^k(x) \in W^s_{loc}$ for sufficiently large k. If $W^s_{loc}(x^*)$ is of measure zero, then $g^{-k}(W^s_{loc}(x^*))$ is also of measure zero, and hence the global stable set is of measure zero. Thus, gradient descent will never converge to x^* from a random initialization.

In Section 4, we formalize the above arguments by showing the existence of an inverse gradient map. The case of degenerate critical points, critical points with zero eigenvalues, is more delicate; the geometry of the global stable set is no longer characterized by only the number of positive eigenvectors. However in Section 4, we show that if a critical point has at least one negative eigenvalue, then the global stable set is of measure zero.

4. Main Results

We now state and prove our main theorem, making our intuition rigorous.

Theorem 4 *Let f be a C^2 function and x^* be a strict saddle. Assume that $0 < \alpha < \frac{1}{L}$, then*

$$\Pr(\lim_k x_k = x^*) = 0.$$

That is, the gradient method never converges to saddle points, provided the step size is not chosen aggressively. Greedy methods that use precise line search may still get stuck at stationary points. However, a short-step gradient method will only converge to minimizers.

Remark 5 *Note that even for the convex functions method, a constant step size slightly less than $1/L$ is a nearly optimal choice. Indeed, for $\theta < 1$, if one runs the gradient method with step size of θ/L on a convex function a convergence rate of $O(\frac{1}{\theta T})$ is attained.*

Remark 6 *When* $\lim_k x_k$ *does not exist, the above theorem is trivially true.*

To prove Theorem 4, our primary tool will be the theory of Invariant Manifolds. Specifically, we will use Stable-Center Manifold theorem developed in Smale (1967); Shub (1987); Hirsch et al. (1977), which allows for a local characterization of the stable set. Recall that a map $g : X \to Y$ is a diffeomorphism if g is a bijection, and g and g^{-1} are continuously differentiable.

Theorem 7 (Theorem III.7, Shub (1987)) *Let* 0 *be a fixed point for the* C^r *local diffeomorphism* $\phi : U \to E$*, where* U *is a neighborhood of* 0 *in the Banach space* E*. Suppose that* $E = E_s \oplus E_u$*, where* E_s *is the span of the eigenvectors corresponding to eigenvalues of magnitude less than or equal to* 1 *of* $D\phi(0)$*, and* E_u *is the span of the eigenvectors corresponding to eigenvalues of magnitude greater than* 1 *of* $D\phi(0)$*. Then there exists a* C^r *embedded disk* W^{cs}_{loc} *that is tangent to* E_s *at* 0 *called the* local stable center manifold*. Moreover, there exists a neighborhood* B *of* 0*, such that* $\phi(W^{cs}_{loc}) \cap B \subset W^{cs}_{loc}$*, and* $\cap_{k=0}^{\infty} \phi^{-k}(B) \subset W^{cs}_{loc}$*.*

To unpack all of this terminology, what the stable manifold theorem says is that if there is a map that diffeomorphically deforms a neighborhood of a critical point, then this implies the existence of a local stable center manifold W^{cs}_{loc} containing the critical point. This manifold has dimension equal to the number of eigenvalues of the Jacobian of the critical point that are less than 1. W^{sc}_{loc} contains all points that are locally forward non-escaping meaning, in a smaller neighborhood B, a point converges to x^* after iterating ϕ only if it is in $W^{cs}_{loc} \cap B$.

Relating this back to the gradient method, replace ϕ with our gradient map g and let x^* be a strict saddle point. We first record a very useful fact:

Proposition 8 *The gradient mapping* g *with step size* $\alpha < \frac{1}{L}$ *is a diffeomorphism.*

We will prove this proposition below. But let us first continue to apply the stable manifold theorem. Note that $Dg(x) = I - \alpha\nabla^2 f(x)$. Thus, the set W^{cs}_{loc} is a manifold of dimension equal to the number of non-negative eigenvalues of the $\nabla^2 f(x)$. Note that by the strict saddle assumption, this manifold has strictly positive codimension and hence has measure zero.

Let B be the neighborhood of x^* promised by the Stable Manifold Theorem. If x converges to x^* under the gradient map, then there exists a T such that $g^t(x) \in B$ for all $t \geq T$. This means that $g^t(x) \in \cap_{k=0}^{\infty} g^{-k}(B)$, and hence, $g^t(x) \in W^{cs}_{loc}$. That is, we have shown that

$$W^s(x^*) \subseteq \bigcup_{l \geq 0}^{\infty} g^{-l}(W^{cs}_{loc}).$$

Since diffeomorphisms map sets of measure zero to sets of measure zero, and countable unions of measure zero sets have measure zero, we conclude that W^s has measure zero. That is, we have proven Theorem 4.

4.1. Proof of Proposition 8

We first check that g is injective from $\mathbf{R}^n \to \mathbf{R}^n$ for $\alpha < \frac{1}{L}$. Suppose that there exist x and y such that $g(x) = g(y)$. Then we would have $x - y = \alpha(\nabla f(x) - \nabla f(y))$ and hence

$$\|x - y\| = \alpha\|\nabla f(x) - \nabla f(y)\| \leq \alpha L\|x - y\|.$$

Since $\alpha L < 1$, this means $x = y$.

To show the gradient map is surjective, we will construct an explicit inverse function. The inverse of the gradient mapping is given by performing the proximal point algorithm on the function $-f$. The proximal point mapping of $-f$ centered at y is given by

$$x_y = \arg\min_x \frac{1}{2} \|x - y\|^2 - \alpha f(x).$$

For $\alpha < \frac{1}{L}$, the function above is strongly convex with respect to x, so there is a unique minimizer. Let x_y be the unique minimizer, then by KKT conditions,

$$y = x_y - \nabla f(x_y) = g(x_y)\,.$$

Hence, x_y is mapped to y by the gradient map.

We have already shown that g is a bijection, and continuously differentiable. Since $Dg(x) = I - \alpha \nabla^2 f(x)$ is invertible for $\alpha < \frac{1}{L}$, the inverse function theorem guarantees g^{-1} is continuously differentiable, completing the proof that g is a diffeomorphism.

4.2. Further consequences of Theorem 4

Corollary 9 *Let C be the set of saddle points and assume they are all strict. If C has at most countably infinite cardinality, then*

$$\Pr(\lim_k x_k \in C) = 0.$$

Proof By applying Corollary 4 to each point $x^* \in C$, we have that $\Pr(\lim_k x_k = x^*) = 0$. Since the critical points are countable, the conclusion follows since countable union of null sets is a null set. ■

Remark 10 *If the saddle points are isolated points, then the set of saddle points is at most countably infinite.*

Theorem 11 *Assume the same conditions as Corollary 9 and $\lim_k x_k$ exists, then $\Pr(\lim_k x_k = x^\star) = 1$, where $x^\star$ is a local minimizer.*

Proof Using the previous theorem, $\Pr(\lim_k x_k \in C) = 0$. Since $\lim_k x_k$ exists and there is zero probability of converging to a saddle, then $\Pr(\lim_k x_k = x^*) = 1$, where x^* is a local minimizer. ■

We now discuss two sufficient conditions for $\lim_k x_k$ to exist. The following proposition prevents x_k from escaping to ∞, by enforcing that f has compact sublevel sets, $\{x : f(x) \leq c\}$. This is true for any coercive function, $\lim_{\|x\| \to \infty} f(x) = \infty$, which holds in most machine learning applications since f is usually a loss function.

Proposition 12 (Proposition 12.4.4 of Lange (2013)) *Assume that f is continuously differentiable, has isolated critical points, and compact sublevel sets, then $\lim_k x_k$ exists and that limit is a critical point of f.*

The second sufficient condition for $\lim_k x_k$ to exist is based on the Lojasiewicz gradient inequality, which characterizes the steepness of the gradient near a critical point. The Lojasiewicz inequality ensures that the length traveled by the iterates of gradient descent is finite. This will also allow us to derive rates of convergence to a local minimum.

Definition 13 (Lojasiewicz Gradient Inequality) *A critical point x^* is satisfies the Lojasiewicz gradient inequality if there exists a neighborhood U, $m, \epsilon > 0$, and $0 \leq a < 1$ such that*

$$\|\nabla f(x)\| \geq m|f(x) - f(x^*)|^a \tag{3}$$

for all x in $\{x \in U : f(x^*) < f(x) < f(x^*) + \epsilon\}$.

The Lojasiewicz inequality is very general as discussed in Bolte et al. (2010); Attouch et al. (2010, 2013). In fact every analytic function satisfies the Lojasiewicz inequality. Also if the solution is μ-strongly convex in a neighborhood, then the Lojasiewicz inequality is satisfied with parameters $a = \frac{1}{2}$, and $m = \sqrt{2\mu}$.

Proposition 14 *Assume the same conditions as Corollary 9, and the iterates do not escape to ∞, meaning $\{x_k\}$ is a bounded sequence. Then $\lim_k x_k$ exists and $\lim_k x_k = x^*$ for a local minimum x^*.*

Furthermore if x^* satisfies the Lojasiewicz gradient inequality for $0 < a \leq \frac{1}{2}$, then for some C and $b < 1$ independent of k,

$$\|x_k - x^*\| \leq Cb^k.$$

For $\frac{1}{2} < a < 1$,

$$\|x_k - x^*\| \leq \frac{C}{k^{(1-a)/(2a-1)}}.$$

Proof The first part of the theorem follows from Absil et al. (2005), which shows that $\lim_k x_k$ exists. By Theorem 11, $\lim_k x_k$ is a local minimizer x^*. Without loss of generality, we may assume that $f(x^*) = 0$ by shifting the function.

Absil et al. (2005) also establish

$$\sum_{j=k}^{\infty} \|x_{j+1} - x_j\| \leq \frac{2}{\alpha m(1-a)} f(x_k)^{1-a}.$$

Define $e_k = \sum_{j=k}^{\infty} \|x_{j+1} - x_j\|$, and since $e_k \geq \|x_k - x^*\|$ it suffices to upper bound e_k.

Since we have established that x_k converges, for k large enough we can use the gradient inequality and $\nabla f(x_k) = \frac{x_k - x_{k+1}}{\alpha}$:

$$\begin{aligned} e_k &\leq \frac{2}{\alpha m(1-a)} f(x_k)^{1-a} \\ &\leq \frac{2}{\alpha m^{1/a}(1-a)} \|\nabla f(x_k)\|^{(1-a)/a} \\ &\leq \frac{2}{(m\alpha)^{1/a}(1-a)} (e_k - e_{k+1})^{(1-a)/a}. \end{aligned}$$

Define $\beta = \frac{2}{(m\alpha)^{1/a}(1-a)}$ and $d = \frac{a}{1-a}$. First consider the case $0 \leq a \leq \frac{1}{2}$, then $d \leq 1$. Thus,

$$\begin{aligned} e_k &\leq \beta(e_k - e_{k+1})^{1/d} \\ e_{k+1} &\leq e_k - \left(\frac{e_k}{\beta}\right)^d \\ &\leq \left(1 - \frac{1}{\beta^d}\right) e_k, \end{aligned}$$

where the last inequality uses $e_k < 1$ and $d \leq 1$.

For $\frac{1}{2} < a < 1$, we have established $e_{k+1} \leq e_k - \left(\frac{e_k}{\beta}\right)^d$. Define the shorthands $t = \frac{1-a}{2a-1}$ and $r = \frac{C^{d-1}}{\beta^d}$. The inductive hypothesis guarantees us $e_k \leq \frac{C}{k^t}$, so $e_{k+1} \leq C\left(\frac{1}{k^t} - \frac{r}{k^{td}}\right)$. We need to verify now that

$$\begin{aligned} \frac{1}{k^t} - \frac{r}{k^{td}} &\leq \frac{1}{(k+1)^t} \\ \frac{(k+1)^t}{k^t} - 1 &\leq r\frac{(k+1)^t}{k^{td}} \end{aligned}$$

The last equation can be explicitly verified by noting that the left-hand side $\frac{(k+1)^t}{k^t} - 1 \leq \frac{2t}{k}$ for k large enough. The right-hand side satisfies $r\frac{(k+1)^t}{k^{td}} \geq r\frac{k^t}{k^{td}} = \frac{r}{k}$. Thus for C large enough, $\frac{2t}{k} \leq \frac{r}{k}$, which completes the proof.

■

5. Conclusion

We have shown that gradient descent with random initialization and appropriate constant step size does not converge to a saddle point. Our analysis relies on a characterization of the local stable set from the theory of invariant manifolds. The geometric characterization is not specific to the gradient descent algorithm. To use Theorem 4, we simply need the update step of the algorithm to be a diffeomorphism. For example if g is the mapping induced by the proximal point algorithm, then g is a diffeomorphism with inverse given by gradient ascent on $-f$. Thus the results in Section 4 also apply to the proximal point algorithm. That is, *the proximal point algorithm does not converge to saddles*. We expect that similar arguments can be used to show ADMM, mirror descent and coordinate descent do not converge to saddle points under appropriate choices of step size. Indeed, convergence to minimizers has been empirically observed for the ADMM algorithm Sun et al. (2015a).

It is not clear if the step size restriction ($\alpha < 1/L$) is necessary to avoid saddle points. Most of the constructions where the gradient method converges to saddle points require fragile initial conditions as discussed in Section 3. It remains a possibility that methods that choose step sizes greedily, by Wolfe Line Search or backtracking, may still avoid saddle points provided the initial point is chosen at random. We leave such investigations for future work.

Another important piece of future work would be relaxing the conditions on isolated saddle points. In a followup work, Panageas and Piliouras (2016) have addressed this by using a countable

covering of $\mathbb{R}^n$, instead of a naive union bound. It is possible that for the structured problems that arise in machine learning, whether in matrix factorization or convolutional neural networks, that saddle points are isolated after taking a quotient with respect to the associated symmetry group of the problem. Techniques from dynamical systems on manifolds may be applicable to understand the behavior of optimization algorithms on problems with a high degree of symmetry.

It is also important to understand how stringent the strict saddle assumption is. Will a perturbation of a function always satisfy the strict saddle property? Adler and Taylor (2009) provide very general sufficient conditions for a random function to be Morse, meaning the eigenvalues at critical points are non-zero, which implies the strict saddle condition. These conditions rely on checking the density of $\nabla^2 f(x)$ has full support conditioned on the event that $\nabla f(x) = 0$. This can be explicitly verified for functions f that arise from learning problems.

However, we note that there are very difficult unconstrained optimization problems where the strict saddle condition fails. Perhaps the simplest is optimization of quartic polynomials. Indeed, checking if 0 is a local minimizer of the quartic

$$f(x) = \sum_{i,j=1}^{n} q_{ij} x_i^2 x_j^2$$

is equivalent to checking whether the matrix $Q = [q_{ij}]$ is co-positive, a co-NP complete problem. For this f, the Hessian at $x = 0$ is zero. In concurrent work, Anandkumar and Ge (2016) have proposed an algorithm to avoid third-order saddles. Interestingly, the strict saddle property failing is analogous in dynamical systems to the existence of a *slow manifold* where complex dynamics may emerge. Slow manifolds give rise to metastability, bifurcation, and other chaotic dynamics, and it would be intriguing to see how the analysis of chaotic systems could be applied to understand the behavior of optimization algorithms around these difficult critical points.

Acknowledgements

The authors would like to thank Chi Jin, Tengyu Ma, Robert Nishihara, Mahdi Soltanolkotabi, Yuekai Sun, Jonathan Taylor, and Yuchen Zhang for their insightful feedback. MS is generously supported by an NSF Graduate Research Fellowship. BR is generously supported by ONR awards N00014-14-1-0024, N00014-15-1-2620, and N00014-13-1-0129, and NSF awards CCF-1148243 and CCF-1217058. MIJ is generously supported by ONR award N00014-11-1-0688 and by the ARL and the ARO under grant number W911NF-11-1-0391. This research is supported in part by NSF CISE Expeditions Award CCF-1139158, DOE Award SN10040 DE-SC0012463, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, IBM, SAP, The Thomas and Stacey Siebel Foundation, Adatao, Adobe, Apple Inc., Blue Goji, Bosch, Cisco, Cray, Cloudera, Ericsson, Facebook, Fujitsu, Guavus, HP, Huawei, Intel, Microsoft, Pivotal, Samsung, Schlumberger, Splunk, State Farm, Virdata and VMware.

References

Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*, 2016.

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

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

Mikhail Belkin, Luis Rademacher, and James Voss. Basis learning as an algorithmic primitive. *arXiv preprint arXiv:1411.1420*, 2014.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, Laurent Mazet, et al. Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc*, 362(6):3319–3363, 2010.

T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *arXiv preprint arXiv:1506.03382*, 2015.

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

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

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

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. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

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

Philip E Gill and Walter Murray. Newton-type methods for unconstrained and linearly constrained optimization. *Mathematical Programming*, 7(1):311–350, 1974.

M.W. Hirsch, C.C. Pugh, and M. Shub. *Invariant Manifolds*. Number no. 583 in Lecture Notes in Mathematics. Springer-Verlag, 1977.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2009.

K Lange. Optimization. springer texts in statistics. 2013.

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.

Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

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

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

Ioannis Panageas and Georgios Piliouras. Gradient descent converges to minimizers: The case of non-isolated critical points. *arXiv preprint arXiv:1605.00405*, 2016.

Razvan Pascanu, Yann N Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization. *arXiv:1405.4604*, 2014.

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

Michael Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 1987.

Stephen Smale. Differentiable dynamical systems. *Bulletin of the American mathematical Society*, 73(6): 747–817, 1967.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *arXiv:1511.03607*, 2015a.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *arXiv:1511.04777*, 2015b.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Forthcoming*, 2016.

Yuchen Zhang, Xi Chen, Denny Zhou, and Michael I Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*, pages 1260–1268, 2014.

Tuo Zhao, Zhaoran Wang, and Han Liu. Nonconvex low rank matrix factorization via inexact first order oracle.