Convergence Results for Gradient Descent

Peter Racioppo

With the increasing need to optimize high-dimensional, non-convex objective functions, studying the structure of these problems, and the efficiency and convergence of first-order methods like gradient descent is a vital area of research. Here, we review three recent papers which make important contributions to these efforts. *Gradient Descent Only Converges to Minimizers* (2016) shows that, under mild assumptions on the objective function, gradient descent always converges to a minimizer. *Gradient Descent Finds Global Minima of Deep Neural Networks* (2019) proves that gradient descent applied to a deep neural network, again under mild assumptions on loss and activation functions, converges to a global minimum during training, provided that gradient descent step sizes are sufficiently small and the network is sufficiently large. Finally, *How to Escape Saddle Points Efficiently* (2017) shows that the efficiency of gradient descent in escaping saddle points can be significantly improved by the addition of appropriate perturbations.

Background (Dynamical Systems Theory):

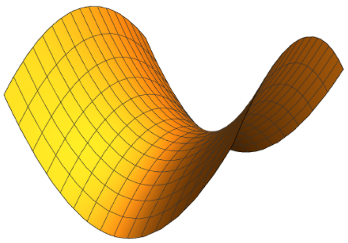
Definition (Fixed Point): We say that $\boldsymbol{x}^*$ is a fixed point of a function $\boldsymbol{f}$ if $\boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{x}^*$. A point $\boldsymbol{x}^*$ is a critical point of $\boldsymbol{f}(\boldsymbol{x})$ if it is a fixed point of the gradient map $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x} - \alpha\nabla\boldsymbol{f}(\boldsymbol{x})$, or equivalently $\nabla\boldsymbol{f}(\boldsymbol{x}^*) = 0$.

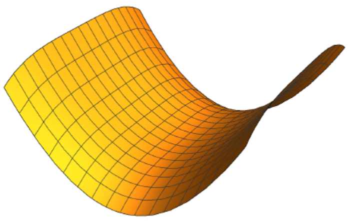
Definition (Isolated Critical Point): A critical point $\boldsymbol{x}^*$ is isolated if there exists a neighborhood of $\boldsymbol{x}^*$ containing no other critical points.

Definition (Minima & Maxima): A critical point is a local minimum if there is a neighborhood U around $\boldsymbol{x}^*$ such that $\boldsymbol{f}(\boldsymbol{x}^*) \leq \boldsymbol{f}(\boldsymbol{x})$ for all $\boldsymbol{x} \in U$, and a local maximum if $\boldsymbol{f}(\boldsymbol{x}^*) \geq \boldsymbol{f}(\boldsymbol{x})$.

Definition (Saddle Point): A critical point is a saddle point if for all neighborhoods U around $\boldsymbol{x}^*$, there are $\boldsymbol{x}, \boldsymbol{y} \in U$ such that $\boldsymbol{f}(\boldsymbol{x}) \leq \boldsymbol{f}(\boldsymbol{x}^*) \leq \boldsymbol{f}(\boldsymbol{y})$.

Definition (Strict Saddle Point): We say that $\boldsymbol{x}^*$ is a strict saddle point if $\nabla^2\boldsymbol{f}(\boldsymbol{x}^*)$ has at least one strictly negative eigenvalue.


Strict saddle point


Non-strict saddle point

Consider the nonlinear system $\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), t),\ \ \boldsymbol{x}_0 = \boldsymbol{x}(t_0).$ (1)

Definition (Lipschitz Continuity): We say that $\boldsymbol{f}(t, \boldsymbol{x})$ is Lipschitz continuous if it satisfies the inequality: $\|\boldsymbol{f}(t, \boldsymbol{x}) - \boldsymbol{f}(t, \boldsymbol{y})\| \leq \boldsymbol{L}\|\boldsymbol{x} - \boldsymbol{y}\|$ for all $(t, \boldsymbol{x})$ and $(t, \boldsymbol{y})$ in some neighborhood of $(t, \boldsymbol{x}_0)$. The Lipschitz condition is stronger than continuity but weaker than continuous differentiability.

Theorem (Local Existence and Uniqueness):
Let $\boldsymbol{f}(t, \boldsymbol{x})$be piecewise continuous in t and satisfy the Lipschitz condition$\|\boldsymbol{f}(t, \boldsymbol{x}) - \boldsymbol{f}(t, \boldsymbol{y})\| \leq \boldsymbol{L}\|\boldsymbol{x} - \boldsymbol{y}\|$ for $\forall\ x, y\ \in B = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq r\}, \forall\ t \in [t_0, t_1]$.
Then, $\exists\ \delta > 0$ such that (1) has a unique solution over $[t_0, t_0 + \delta]$.

Theorem (Global Existence and Uniqueness):
If $\boldsymbol{f}(t, \boldsymbol{x})$ satisfies piecewise continuity and the Lipschitz condition for $\forall\ x, y\ \in \mathbb{R}^n, \forall\ t \in [t_0, t_1]$, then (1) has a unique solution on $[t_0, t_1]$.

Definition (Equilibrium Point): We say that the point $\boldsymbol{x}^*$ is an equilibrium point of (1) if $\boldsymbol{f}(0, \boldsymbol{x}^*) = 0\ \Rightarrow \boldsymbol{f}(t, \boldsymbol{x}^*) = 0, \forall\ t \geq 0$. (Whenever the system starts at $\boldsymbol{x}^*$, it remains at $\boldsymbol{x}^*$ for all future time.)

Definition (Stability):
Consider the nonlinear system (1) where $\boldsymbol{f}(t, \boldsymbol{x})$ is piecewise continuous in t and locally Lipschitz in $\boldsymbol{x}$. Let $\boldsymbol{x} = 0$ be an equilibrium point. (We can always take the equilibrium point to be at the origin, because any equilibrium point can be shifted to the origin by a change of variables.)

The equilibrium point is:

- Stable if, for each $\epsilon > 0,\ \exists\ \delta(\epsilon, t_0)$ such that $\|\boldsymbol{x}(t_0)\| < \delta \Rightarrow\ \|\boldsymbol{x}(t)\| < \epsilon$, for $\forall\ t \geq t_0 \geq 0$.
- Uniformly stable if, for each $\epsilon > 0$ in the previous line, the choice of δ can be made independent of t_0.
- Asymptotically stable if it is stable and $\delta(t_0)$ can be chosen such that $\|\boldsymbol{x}(t_0)\| < \delta \Rightarrow \lim_{t\to\infty} \boldsymbol{x}(t) = 0$.
- Unstable if not stable.

Theorem (Stability of a Linear Time Invariant (LTI) System): Consider the linear time invariant system $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$. The equilibrium point $\boldsymbol{x} = 0$ is (globally) asymptotically stable if and only if all eigenvalues of $\boldsymbol{A}$ have strictly negative real part.

Theorem (Stability of a Discrete-Time LTI System): Consider the linear difference equation $\boldsymbol{x}(t + 1) = \boldsymbol{A}\boldsymbol{x}(t)$. The unique equilibrium point is asymptotically stable if and only if all eigenvalues of $\boldsymbol{A}$ have magnitude strictly less than 1.

Definition (Stable/Unstable Subspaces): The span of the eigenvectors corresponding to eigenvalues with negative real-part is called the stable subspace E^s, and the subspace spanned by the eigenvectors of eigenvalues with positive real-part is called the unstable subspace E^u. The subspace spanned by the eigenvectors belonging to eigenvalues with zero real-part is called the center subspace E^c.

Theorem (Lyapunov's Indirect Method): Let $\boldsymbol{x}_0 = 0$ be an equilibrium point of $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ (i.e. $f(\boldsymbol{x}_0) = 0$), where $\boldsymbol{f}$ is continuously differentiable. Linearizing about $\boldsymbol{x}_0$, $\dot{\boldsymbol{x}} \approx \boldsymbol{f}(\boldsymbol{x}_0) + \nabla\boldsymbol{f}(\boldsymbol{x})|_{x_0}(\boldsymbol{x} - \boldsymbol{x}_0)$, so $\dot{\boldsymbol{x}} \approx \nabla\boldsymbol{f}(\boldsymbol{x})|_{x_0}(\boldsymbol{x})$. The point $\boldsymbol{x}_0$ is asymptotically stable if all eigenvalues of $\nabla\boldsymbol{f}(\boldsymbol{x})|_{x_0}$ have strictly negative real part.

Definition (Flow of an ODE): The flow of (1) is defined as $\varphi^{t,t_0}(\boldsymbol{x}_0) = \boldsymbol{x}(t + t_0)$. Informally, the flow is the (continuous) trajectory (solution) of the ODE.

Definition (Differentiable Manifold): For our purposes, think of a *k*-dimensional manifold as the solution of the equation $\boldsymbol{\eta}(\boldsymbol{x}) = 0$ with $\boldsymbol{\eta}$: $\mathbb{R}^n \to \mathbb{R}^{n-k}$ satisfying some smoothness conditions (that is, sufficiently many times differentiable): in other words, a *k*-dimensional surface in an *n*-dimensional space. A manifold is said to be invariant if $\boldsymbol{\eta}(\boldsymbol{x}(0)) = 0 \Rightarrow \boldsymbol{\eta}(\boldsymbol{x}(t)) = 0, \forall\ t\ \in [0, t_1)$.

Theorem (Stable Manifold Theorem): Let $\boldsymbol{x}_0 = 0$ be an equilibrium point of $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$, let *E* be an open subset of $\mathbb{R}^n$ containing the origin, and let $\boldsymbol{f}$ be a continuously differentiable function in *E*. Let $\boldsymbol{\phi}^t$ be the flow (let t_0= 0). Suppose that $\nabla\boldsymbol{f}(\boldsymbol{x})|_{x=0}$ has k eigenvalues with negative real part and *n* − *k* eigenvalues with positive real part. Then there exists a *k*-dimensional differentiable manifold *S* tangent to the stable subspace E^s of the linearized system $\dot{\boldsymbol{x}} = \nabla\boldsymbol{f}(\boldsymbol{x})|_{x_0}(\boldsymbol{x})$ such that *S* is invariant and for all $\boldsymbol{x}_0 \in S$, $\lim_{t\to\infty} \varphi^t(\boldsymbol{x}_0) = 0$; and there exists an *n* − *k* differentiable manifold *U* tangent to the unstable subspace E^u such that *U* is invariant and for all $\boldsymbol{x}_0 \in U$, $\lim_{t\to-\infty} \varphi^t(\boldsymbol{x}_0) = 0$.

Gradient Descent Only Converges to Minimizers [1]

(Lee , Simchowitz , Jordan, & Recht)

The presence of saddle-points poses a potential difficulty in non-convex optimization problems. Many examples exist in which finding a local minimum is NP-Hard, and in which worst-case initialized gradient descent provably converges to saddle points. This difficulty is compounded by the fact that most natural objective functions have exponentially more saddle points than minima. Past solutions to this difficulty have relied on second-order methods, but these methods are 2nd or 3rd order polynomial in the dimension of the problem, so the effectiveness of first-order methods is of great importance in high-dimensional settings. The authors of the present paper use the Stable Manifold Theorem to show that, under the assumption that the objective function has only strict saddle points and some mild smoothness conditions, and given a small enough step size, gradient descent with random initialization never converges to saddle points. They then show that, provided that $\lim_{k\to\infty} x^k$ exists, gradient descent converges to a local minimizer with probability 1. Finally, the authors discuss two different sufficient conditions for $\lim_{k\to\infty} x^k$ to exist. The strict saddle point assumption turns out to be applicable in a wide variety of common optimization problems, including "PCA, a fourth-order tensor factorization, formulations of dictionary learning, and phase retrieval."

Definition (Diffeomorphism): A map g: $X \rightarrow Y$ is a diffeomorphism if g is a bijection, and g and g^{-1} are continuously differentiable.

Intuition:

Consider a non-convex quadratic form $\boldsymbol{f}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x}$.
Applying gradient descent: $x_{k+1} = x_k - \alpha\nabla f(x_k) = (I - \alpha H)x_k = (I - \alpha H)^{k+1}x_0$.
Assume, without loss of generality, that $H = \text{diag}(\lambda_1, \dots, \lambda_n)$, with $\lambda_1, \dots, \lambda_p > 0$ and $\lambda_{p+1}, \dots, \lambda_n < 0$. Then, $x_{k+1} = \sum_{i=1}^{n}(1 - \alpha\lambda_i)^{k+1}\langle e_i, x_0\rangle e_i$.
Suppose $\alpha < 1/L$, with $L = \max\ |\lambda_i|$. Note that $(1 - \alpha\lambda_i) < 1$ for $i \leq p$ and $(1 - \alpha\lambda_i) > 1$ for $i > p$. Thus, $\lim_{k\to\infty}(1 - \alpha\lambda_i)^{k+1} = \ 0$ for $\forall\ 1 \leq i \leq p$.
If $x_0 \in E_s = \text{span}(e_1, \dots, e_p)$, then $(1 - \alpha\lambda_i)^{k+1}\langle e_i, x_0\rangle e_i < 0$ for $\forall\ 1 \leq i \leq p$ and $\langle e_i, x_0\rangle = 0$ for $\forall\ p + 1 \leq i \leq n$, so that $\lim_{k\to\infty} x_{k+1} = \ 0$, which is the unique critical point (a saddle point).
On the other hand, if $x_0 \in \text{span}(e_{p+1}, \dots, e_n)$, then $(1 - \alpha\lambda_i)^{k+1}\langle e_i, x_0\rangle e_i > 0$ for some $i \in \mathbb{Z} \cap [p + 1, n]$. Thus, $\lim_{k\to\infty} x_{k+1} = \infty$.
Intuitively, $\text{span}(e_1, \dots, e_p)$ is a p-dimensional space embedded in the n-dimensional space $\text{span}(e_1, \dots, e_n)$. The chance of randomly selecting a vector x_0 that lies completely in the lower dimensional space is zero. In other words, the components of x_0 for indices $i \in \mathbb{Z} \cap [p + 1, n]$ must all be zero, and the probability of randomly selecting a point z on the real line is $\text{P}(\{z\}) = 1 - \text{P}(\{z\}^C) = 1 - 1 = 0$. Thus, the probability of converging to the saddle point is zero.

Main Result:

Theorem: The gradient mapping g with step size $\alpha < 1/L$ is a diffeomorphism.
The authors give another statement of the Stable Manifold Theorem, using differential topology:

Theorem (Stable Manifold Theorem): Let $\boldsymbol{x}_0 = 0$ be a fixed point for the r-continuously differentiable local diffeomorphism ϕ. Let E^s be the span of the eigenvectors corresponding to the stable eigenvalues of $\text{D}\phi(0)$ (D denoting the derivative on the manifold). Then there exists an r-continuously differentiable manifold W_{loc}^{cs} tangent to the stable subspace E^s at 0 that is invariant with respect to ϕ, or as the authors put it: there exists a neighborhood B of 0 such that $\phi(W_{loc}^{cs}) \cap B \subset W_{loc}^{cs}$ and $\bigcap_{k=0}^{\infty}\phi^{-k}(B) \subset W_{loc}^{cs}$. (In the case $\phi = g$, we have that $\boldsymbol{Dg}(\boldsymbol{x}) = I - \alpha\nabla^2\boldsymbol{f}$, so that the dimension of W_{loc}^{cs} is equal to the number of nonnegative eigenvalues of $\nabla^2\boldsymbol{f}$ (I think there's a requirement here that α be sufficiently small).)

Theorem (Gradient Descent Never Converges to Saddle Points): Let f be a real-valued, twice-continuously differentiable function with Lipschitz gradient with constant L. Let x^* be a strict saddle. Assume that $0 < \alpha < \frac{1}{L}$. Then $\Pr\left(\lim_{k\to\infty}\boldsymbol{x}^k = \boldsymbol{x}^*\right) = 0$. In other words, given a sufficiently small step size, gradient descent never converges to saddle points.

Proof: "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_{loc}^{cs} has measure zero." In other words, W_{loc}^{cs} is a null set.

Corollary: Let C be the set of saddle points and assume they are all strict. If C has at most countably infinite cardinality, then $\Pr\left(\lim_{k\to\infty} \boldsymbol{x}^k \in C\right) = 0$. (Note that if the saddle points are isolated points, then the set of saddle points is at most countably infinite.)

Theorem: Provided that the assumptions of the previous theorem hold and that $\lim_{k\to\infty} x^k$ exists, there is zero probability of converging to a saddle, which implies that $\Pr\left(\lim_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^*\right) = 1$.

The remainder of the paper discusses sufficient conditions on $\boldsymbol{f}$ that guarantee the existence of $\lim_{k\to\infty} x^k$. In the case that the critical point $\boldsymbol{x}^*$ of $\boldsymbol{f}$ satisfies the *Lojasiewicz gradient inequality*, the authors derive bounds on the rates of convergence to a local minimum.

Gradient Descent Finds Global Minima of Deep Neural Networks [2]

(Du, Lee, Li, Wang, Zhai)

With the increasing ubiquity of deep learning, much recent work has focused on the optimization of neural networks, a generally non-convex optimization problem. One focus of study has been the set of functions which has only strict saddle points (that is, saddle points with strictly negative curvature in at least one direction). However, a non-strict saddle point exists in even the three-layer case. In contact, the approach employed in the present paper is to study the training loss. Recent papers from 2017-2018 have shown global convergence of gradient descent in some shallow networks, and have derived training convergence rates in e.g. two-layer feedforward networks and recurrent neural networks. The present paper extends these results to deeper networks and to ResNet and convolutional architectures.

In sufficiently large neural networks, randomly-initialized first-order methods such as gradient descent achieve zero training loss. In a fully-connected feedforward network with a smooth, Lipschitz-continuous activation function at each neuron, this paper proves an upper bound on the network size, such that randomly-initialized gradient descent converges to zero training loss. In particular, in a feedforward network with H layers of width m, $m = \Omega(poly(n)2^{O(h)})$ is such an upper bound.

Outline of the Proof Technique: Let $\boldsymbol{y}$ be the vector of labels and $\boldsymbol{u}$ the vector of predicted labels at the kth iteration, whose ith component is $u_i(k) = f(\theta(k), \boldsymbol{x}_i)$, where θ is the vector of network parameters and $\boldsymbol{x}$ is the input. Define the Gram matrix at the kth iteration $\mathbf{G}(k)$, with ijth component $\mathbf{G}_{ij}(k) = \langle \frac{\partial u_i(k)}{\partial \theta(k)}, \frac{\partial u_j(k)}{\partial \theta(k)} \rangle$. Du et al. showed in a 2018 paper that, for a 2-layer fully-connected neural network with gradient descent step-size η, the difference between predicted and actual labels obeys the dynamics $\boldsymbol{y} - \boldsymbol{u}(k+1) = (I - \eta\mathbf{G}(k))(\boldsymbol{y} - \boldsymbol{u}(k))$, and that $\mathbf{G}(k)$ is approximately equal to a constant matrix $\mathbf{K}$ for large m. The dynamics of $\{\boldsymbol{y} - \boldsymbol{u}(k+1)\}_{k=0}^{\infty}$ are then linear, and standard power method techniques show that the sequence $\{\boldsymbol{y} - \boldsymbol{u}(k+1)\}_{k=0}^{\infty}$ converges to 0. Extending this result to *deep* neural networks requires more sophisticated techniques and a two-part proof. The authors first show that the Gram matrix at the Hth-layer in the initialization phase $\mathbf{G}^{(H)}(0)$ is close to a constant matrix $\mathrm{K}^{(H)}$, and then show that that the Gram matrix during the kth training step $\mathbf{G}^{(H)}(k)$ is close to $\mathbf{G}^{(H)}(0)$ for all k. The authors also prove that

$\mathbf{K}^{(H)}$ is strictly positive definite, provided the training data is nondegenerate, and thus, from the analysis of the power method, that gradient descent has a linear rate of convergence. In a fully-connected network, the upper bound on m has exponential dependency on H, but in the case of the ResNet architecture, the authors show that this can be reduced to a polynomial dependency.

Citations

[1] J. Lee , M. Simchowitz , M. Jordan, B. Recht. *Gradient Descent Only Converges to Minimizers.* 2016. JMLR: Workshop and Conference Proceedings vol 49:1–12, 2016.

[2] S. Du, J. Lee, H. Li, L. Wang, X. Zhai. *Gradient Descent Finds Global Minima of Deep Neural Networks*. 2019. arXiv:1811.03804

[3] C. Jin, R. Ge, P. Netrapalli, S. Kakade, M. Jordan. *How to Escape Saddle Points Efficiently*. 2017. arXiv:1703.00887.

[4] H. Khalil. *Nonlinear Systems, 3ed*. Prentice Hall. 2002.