Journal of Machine Learning Research 1 (2017) Submitted ; Published

Representing smooth functions as compositions of near-identity functions with implications for deep network optimization

Peter L. Bartlett PETER@BERKELEY.EDU
Steven N. Evans EVANS@STAT.BERKELEY.EDU
Philip M. Long PLONG@GOOGLE.COM

Editor:

Abstract

We show that any smooth bi-Lipschitz h can be represented exactly as a composition $h_m \circ ... \circ h_1$ of functions $h_1, ..., h_m$ that are close to the identity in the sense that each $(h_i - \mathrm{Id})$ is Lipschitz, and the Lipschitz constant decreases inversely with the number m of functions composed. This implies that h can be represented to any accuracy by a deep residual network whose nonlinear layers compute functions with a small Lipschitz constant. Next, we consider nonlinear regression with a composition of near-identity nonlinear maps. We show that, regarding Fréchet derivatives with respect to the $h_1, ..., h_m$, any critical point of a quadratic criterion in this near-identity region must be a global minimizer. In contrast, if we consider derivatives with respect to parameters of a fixed-size residual network with sigmoid activation functions, we show that there are near-identity critical points that are suboptimal, even in the realizable case. Informally, this means that functional gradient methods for residual networks cannot get stuck at suboptimal critical points corresponding to near-identity layers, whereas parametric gradient methods for sigmoidal residual networks suffer from suboptimal critical points in the near-identity region.

Keywords: Deep learning, residual networks, optimization.

1. Introduction

The winner of the ILSVRC 2015 classification competition used a new architecture called residual networks (He et al., 2016), which enabled very fast training of very deep networks. These have since been widely adopted. (As of this writing, the paper that introduced this technique, published in 2015, has over 3700 citations.) Deep networks express models as the composition of transformations; residual networks depart from traditional deep learning models by using parameters to describe how each transformation differs from the identity, rather than how it differs from zero.

Motivated by this methodological advance, Hardt and Ma (2017) recently considered compositions of many linear maps, each close to the identity map. They showed that any matrix with spectral norm and condition number bounded by constants can be represented as a product of matrices $I + A_i$, where each A_i has spectral norm $O(1/d)$. They considered this non-convex parameterization for a linear regression problem with additive Gaussian noise, and showed that any critical point of the quadratic loss for which the A_i have sufficiently small spectral norm must correspond to the linear transformation that generated the data. This raised the possibility that gradient descent with each layer initialized to the identity might provably converge for this non-convex optimization problem;

Bartlett et al. (2018) investigated this, identifying sets of problems where this method converges, and where it does not.

In this paper, we continue this line of research. First, we identify a non-linear counterpart of Hardt and Ma's results motivated by deep residual networks: any smooth bi-Lipschitz h (that is, invertible Lipschitz map with differentiable inverse) can be represented exactly as a composition of functions h_i that are close to the identity in the sense that each $(h_i - \mathrm{Id})$ is Lipschitz, and the Lipschitz constant decreases inversely with the number of functions composed. Since a two-layer neural network with standard activation functions can approximate arbitrary continuous functions, we can represent each h_i in the composition as $h_i = \mathrm{Id} + N_i$, where N_i is computed by a two-layer network in this way. The fact that $(h_i - \mathrm{Id})$ has a small Lipschitz constant for deep networks shows that N_i is small, in the sense that it only needs to approximate a slowly changing function.

The requirement in our analysis that h is bi-Lipschitz generalizes the assumption in the linear case studied by Hardt and Ma that the map to be learned has a bounded condition number. The practical strength, and therefore relevance, of invertible feature maps in the non-linear case is supported by success the reversible networks (Maclaurin et al., 2015; Gomez et al., 2017; **?**).

For our second result, we consider a nonlinear regression problem using a composition of near-identity nonlinear maps. If we consider Fréchet derivatives with respect to the functions h_i in the composition, we show that any critical point of the quadratic criterion must be the global optimum. In contrast, if each h_i is a two-layer net of the form $A_i \tanh(B_i x) + x$, analogously to the architecture of (He et al., 2016), and we consider derivatives with respect to the real-valued parameters, there are regression problems that give rise to suboptimal critical points. We discuss the implications of this analysis in Section 6.

A number of authors have investigated how using deep architecture affects the set of functions computed by a network (see Montufar et al., 2014; Telgarsky, 2015; Poole et al., 2016; Mhaskar et al., 2016). Our main results abstract away the parameterization and focus on the expressiveness and of compositions of near-identity functions, along with properties of their error landscapes.

2. Notation and Definitions

Let Id denote the identity map on $\mathbb{R}^d$, $\mathrm{Id}(x) = x$. Throughout, $\|\cdot\|$ denotes a norm on $\mathbb{R}^d$. We also use $\|\cdot\|$ to denote an induced norm: for a function $f : U \to V$, where U and V are normed spaces with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, we write $\|f\| := \sup\left\{\frac{\|f(x)\|_V}{\|x\|_U} : x \in U,\, \|x\|_U > 0\right\}$. Define the Lipschitz seminorm of f as

$$\|f\|_L := \sup\left\{\frac{\|f(x) - f(y)\|_V}{\|x - y\|_U} : x, y \in U,\, x \neq y\right\}.$$

Define the ball of radius α in a normed space $(\mathcal{X}, \|\cdot\|)$ as $B_\alpha(\mathcal{X}) = \{x \in \mathcal{X} : \|x\| \leq \alpha\}$. For a function $f : \mathbb{R}^d \to \mathbb{R}^d$, Df denotes the Jacobian matrix, that is, the matrix with entries $J_{i,j}(x) = \partial f_i / \partial x_j(x)$.

For a functional $F : U \to V$ defined on Banach spaces U and V, recall that the Fréchet derivative of F at $f \in U$ is the linear operator $DF(f) : U \to V$ satisfying $DF(f)(\Delta) = F(f + \Delta) - F(f) + o(\Delta)$, that is,

$$\lim_{\Delta \to 0} \frac{\|F(f + \Delta) - F(f) - DF(f)(\Delta)\|_V}{\|\Delta\|_U} = 0.$$

We use $D_f G(f,g)$ to denote the Fréchet derivative of $f \mapsto G(f,g)$.

3. Representation

Theorem 1 *For $R > 0$, denote $\mathcal{X} = B_R(\mathbb{R}^d)$. Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable, invertible map satisfying the following properties: (a) Smoothness: for some $\alpha > 0$ and all $x, y, u \in \mathcal{X}$,*

$$\|(Dh(y) - Dh(x))\, u\| \leq \alpha \|y - x\| \|u\|; \tag{1}$$

(b) Lipschitz inverse: for some $M > 0$, $\|h^{-1}\|_L \leq M$; (c) Positive orientation: For some $x_0 \in \mathcal{X}$, $\det(Dh(x_0)) > 0$.

Then for all m, there are m functions $h_1, \ldots, h_m : \mathbb{R}^d \to \mathbb{R}^d$ satisfying, for all $x \in \mathcal{X}$, $h_m \circ h_{m-1} \circ \cdots \circ h_1(x) = h(x)$ and, on $h_{i-1} \circ \cdots \circ h_1(\mathcal{X})$, $\|h_i - \mathrm{Id}\|_L = O(\log m/m)$.

Think of the functions h_i as near-identity maps that might be computed as $h_i(x) = x + A\sigma(Bx) + b$, where $A \in \mathbb{R}^{d \times k}$ and $B \in \mathbb{R}^{k \times d}$ are matrices, σ is a nonconstant nonlinearity, such as a sigmoidal function or piecewise-linear function, applied component-wise, and $b \in \mathbb{R}^d$ is a vector. Although the proof constructs the h_i as differentiable (and even smooth) maps, each could, for example, be approximated to arbitrary accuracy on the compact $\mathcal{X}$ using a single layer of ReLUs (for which $\sigma(\alpha) = \max\{0, \alpha\}$). (See, for example, Theorem 1 in (Hornik, 1991), and the comments in Section 3 of that paper about immediate generalizations to unbounded nonlinearities.) In that case, the conclusion of the theorem implies that $x \mapsto A\sigma(Bx)$ can be $O(\log m/m)$-Lipschitz.

Notice that the conclusion of the theorem does not require the function $(h_i - \mathrm{Id})$ to be small; shifting h_i by an arbitrary constant does not affect the Lipschitz property.

The constants hidden in the big-oh notation in the theorem are polynomial in $1/\alpha$, R, M, $|\log \sigma_{\max}(Dh(x_0))|$, and $|\log \sigma_{\min}(Dh(x_0))|$. (Here, $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values.)

The condition $\det(D(h(x_0))) > 0$ is an unavoidable topological constraint that arises because of the orientation of the identity map. As Hardt and Ma (2017) argue in the linear context, if we view h as a mapping from raw representations to meaningful features, we can easily set the orientation of h appropriately (that is, so that $\det(D(h(x_0))) > 0$) without compromising the mapping's usefulness.

To prove Theorem 1, we prove the following special case.

Theorem 2 *Consider an h that satisfies the conditions of Theorem 1 and also $h(0) = 0$ and $Dh(0) = I$. Then for all m, there are m functions $h_1, \ldots, h_m : \mathbb{R}^d \to \mathbb{R}^d$ satisfying, for all $x \in \mathcal{X}$,*

$$h_m \circ h_{m-1} \circ \cdots \circ h_1(x) = h(x)$$

and, on $h_{i-1} \circ \cdots \circ h_1(\mathcal{X})$, $\|h_i - \mathrm{Id}\|_L \leq \epsilon$, provided $\epsilon \geq \frac{B \ln 2m}{m-1}$, where the constant B depends on α, R and M.

To see that Theorem 1 is a corollary, notice that we can write $h(x) = Dh(x_0)\tilde{h}(x - x_0) + h(x_0)$, where $\tilde{h}(x) := (Dh(x_0))^{-1} (h(x + x_0) - h(x_0))$. Since $\tilde{h}$ satisfies $\tilde{h}(0) = 0$ and $D\tilde{h}(0) = I$, Theorem 2 shows that it can be expressed as a composition of near-identity maps. Furthermore, the translations $t_{x_0}(x) := x - x_0$ and $t_{h(x_0)}(x) := x - h(x_0)$ have the property that $(t_{x_0} - \mathrm{Id})$ is 0-Lipschitz. Finally, Theorem 2.1 in Hardt and Ma (2017) shows that we can decompose the Jacobian matrix $Dh(x_0) = (I + A_1) \cdots (I + A_m)$ with $\|A_i\| = O(\gamma/m)$ for $\gamma = |\log \sigma_{\max}(Dh(x_0))| +$

$|\log \sigma_{\min}(Dh(x_0))|$, and this implies that the linear map A_i is $O(\gamma/m)$-Lipschitz (see Lemma 5, Part 3 below).

Before proving Theorem 2, we observe that the smoothness property implies a bound on the accuracy of a linear approximation, and a Lipschitz bound. The proof is in Appendix A.

Lemma 3 *For h satisfying the conditions of Theorem 2 and any $x, y \in \mathcal{X}$,*

$$\|h(y) - (h(x) + Dh(x)(y - x))\| \leq \frac{\alpha}{2}\|y - x\|^2,$$

and $\|h\|_L \leq 1 + \alpha R$.

Proof (of Theorem 2) We give an explicit construction of the $h_1, \ldots, h_m$. For $i = 1, \ldots, m$, define $g_i : \mathcal{X} \to \mathbb{R}^d$ by $g_i(x) = h(a_i x)/a_i$, where the constants $0 < a_1 < \cdots < a_m = 1$ will be chosen later. The g_i's can be viewed as functions that interpolate between the identity (which is $Dh(0)$, the limit as a approaches zero of $h(ax)/a$) and h (which is g_m, because $a_m = 1$). Note that g_i is invertible on $\mathcal{X}$, with $g_i^{-1}(y) = h^{-1}(a_i y)/a_i$ for $y \in g_i(\mathcal{X})$. Define $h_1 = g_1$ and, for $1 < i \leq m$, define $h_i : g_{i-1}(\mathcal{X}) \to \mathbb{R}^d$ by $h_i(x) = g_i(g_{i-1}^{-1}(x))$, so that $h_i \circ h_{i-1} \circ \cdots \circ h_1 = g_i$ and in particular $h_m \circ \cdots \circ h_1 = g_m = h$. It remains to show that, for a suitable choice of $a_1, \ldots, a_{m-1}$, the h_i satisfy the Lipschitz condition.

We have

$$\begin{aligned}
&\|h_1(x) - x - (h_1(y) - y)\| \\
&= \frac{1}{a_1}\|h(a_1 x) - a_1 x - (h(a_1 y) - a_1 y)\| \\
&= \frac{1}{a_1}\|Dh(a_1 y)(a_1 x - a_1 y) - (a_1 x - a_1 y) + (h(a_1 x) - (h(a_1 y) + Dh(a_1 y)(a_1 x - a_1 y)))\| \\
&= \frac{1}{a_1}\|(Dh(a_1 y) - Dh(0))(a_1 x - a_1 y) + (h(a_1 x) - (h(a_1 y) + Dh(a_1 y)(a_1 x - a_1 y)))\| \\
&\leq a_1 \alpha \left(\|y\|\|x - y\| + \frac{1}{2}\|x - y\|^2\right) \quad \text{(by (1) and Lemma 3)} \\
&\leq 2R a_1 \alpha \|x - y\|.
\end{aligned}$$

Now, fix $i > 1$ and $u, v \in \mathcal{X}$ and set $x = g_{i-1}(u)$ and $y = g_{i-1}(v)$. Then we have

$$\begin{aligned}
\|h_i(y) - y - (h_i(x) - x)\| &= \|g_i(v) - g_{i-1}(v) - (g_i(u) - g_{i-1}(u))\| \\
&= \frac{1}{a_i}\left\|h(a_i v) - \frac{a_i}{a_{i-1}}h(a_{i-1}v) - \left(h(a_i u) - \frac{a_i}{a_{i-1}}h(a_{i-1}u)\right)\right\|. \quad (2)
\end{aligned}$$

We consider two cases: when y and x are close, and when they are distant. First, suppose that $\|y-x\| \leq a_i - a_{i-1}$.

$$
\begin{aligned}
&\|h_i(y) - y - (h_i(x) - x)\| \\
&= \frac{1}{a_i}\left\| h(a_i v) - h(a_i u) - \frac{a_i}{a_{i-1}}\left(h(a_{i-1}v) - h(a_{i-1}u)\right)\right\| \\
&= \frac{1}{a_i}\Bigg\| a_i Dh(a_i u)(v-u) + h(a_i v) - (h(a_i u) + a_i Dh(a_i u)(v-u)) \\
&\qquad - \frac{a_i}{a_{i-1}}\left(a_{i-1}Dh(a_{i-1}u)(v-u) + h(a_{i-1}v) - (h(a_{i-1}u) + a_{i-1}Dh(a_{i-1}u)(v-u))\right)\Bigg\| \\
&\leq \|\left(Dh(a_i u) - Dh(a_{i-1}u)\right)(v-u)\| + \frac{\alpha}{2}(a_i + a_{i-1})\|v-u\|^2 \qquad \text{(applying Lemma 3 twice)} \\
&\leq \alpha(a_i - a_{i-1})\|u\|\|v-u\| + \frac{\alpha}{2}(a_i + a_{i-1})\|v-u\|^2 \qquad \text{(by (1))} \\
&\leq \alpha\left(R(a_i - a_{i-1}) + \frac{1}{2}(a_i + a_{i-1})\|v-u\|\right)\|v-u\|.
\end{aligned}
$$

Also, we can relate $\|v-u\|$ to $\|y-x\|$ via the Lipschitz property of h^{-1}:

$$
a_{i-1}\|v-u\| = \|h^{-1}(h(a_{i-1}v)) - h^{-1}(h(a_{i-1}u))\| \leq M\|h(a_{i-1}v) - h(a_{i-1}u)\|,
$$

so

$$
\|y-x\| = \frac{1}{a_{i-1}}\|h(a_{i-1}v) - h(a_{i-1}u)\| \geq \frac{1}{M}\|v-u\|. \tag{3}
$$

Combining, and using the assumption $\|y-x\| \leq a_i - a_{i-1}$,

$$
\begin{aligned}
\|h_i(y) - y - (h_i(x) - x)\| &\leq \alpha M\left(R(a_i - a_{i-1}) + \frac{1}{2}(a_i + a_{i-1})M\|y-x\|\right)\|y-x\| \\
&\leq (a_i - a_{i-1})\alpha M\left(R + M\right)\|y-x\|. \qquad (4)
\end{aligned}
$$

Now suppose that $\|y-x\| > a_i - a_{i-1}$. From (2), we have

$$
\begin{aligned}
&\|h_i(y) - y - (h_i(x) - x)\| \\
&= \frac{1}{a_i}\left\| h(a_i v) - \frac{a_i}{a_{i-1}}h(a_{i-1}v) - \left(h(a_i u) - \frac{a_i}{a_{i-1}}h(a_{i-1}u)\right)\right\| \\
&= \frac{1}{a_i}\left\| \frac{a_{i-1} - a_i}{a_{i-1}}(h(a_{i-1}v) - h(a_{i-1}u)) + h(a_i v) - h(a_{i-1}v) - (h(a_i u) - h(a_{i-1}u))\right\| \\
&= \frac{1}{a_i}\Bigg\| \frac{a_{i-1} - a_i}{a_{i-1}}(h(a_{i-1}v) - h(a_{i-1}u)) \\
&\qquad + h(a_i v) - (h(a_{i-1}v) + Dh(a_{i-1}v)(a_i v - a_{i-1}v)) \\
&\qquad - (h(a_i u) - (h(a_{i-1}u) + Dh(a_{i-1}u)(a_i u - a_{i-1}u))) \\
&\qquad - Dh(a_{i-1}v)(a_i v - a_{i-1}v) + Dh(a_{i-1}u)(a_i u - a_{i-1}u)\Bigg\| \\
&\leq \frac{a_i - a_{i-1}}{a_i}L\|v-u\| + \frac{1}{a_i}\|Dh(a_{i-1}v)(a_i v - a_{i-1}v) - Dh(a_{i-1}u)(a_i u - a_{i-1}u)\| \\
&\qquad + \frac{1}{a_i}\frac{\alpha}{2}(a_i - a_{i-1})^2\left(\|v\|^2 + \|u\|^2\right) \qquad (5)
\end{aligned}
$$

where, in the first term, we have used the Lipschitz property from Lemma 3, with $L = (1+\alpha R)$. But

$$\begin{aligned}
&\frac{1}{a_i}\left\|Dh(a_{i-1}v)(a_iv - a_{i-1}v) - Dh(a_{i-1}u)(a_iu - a_{i-1}u)\right\| \\
&= \frac{a_i - a_{i-1}}{a_i}\left\|Dh(a_{i-1}v)(v) - Dh(a_{i-1}u)(u)\right\| \\
&= \frac{a_i - a_{i-1}}{a_i}\left\|v - u + \left(Dh(a_{i-1}u) - Dh(0)\right)(v-u) + \left(Dh(a_{i-1}v) - Dh(a_{i-1}u)\right)v\right\| \\
&\le \frac{a_i - a_{i-1}}{a_i}\left(1 + \alpha a_{i-1}\left(\|u\| + \|v\|\right)\right)\|v-u\|,
\end{aligned}$$

by (1). Substituting into (5), and using (3) together with the assumption that $\|y - x\| > a_i - a_{i-1}$,

$$\begin{aligned}
&\|h_i(y) - y - (h_i(x) - x)\| \\
&\le \frac{a_i - a_{i-1}}{a_i}\left(LM + M\left(1 + a_{i-1}\alpha(\|u\| + \|v\|)\right) + \frac{\alpha}{2}\left(\|v\|^2 + \|u\|^2\right)\right)\|y - x\| \\
&\le \frac{a_i - a_{i-1}}{a_i}\left(M(L + 1 + 2R\alpha) + \alpha R^2\right)\|y - x\|.
\end{aligned}$$

Combining with (4), it suffices to choose a_1 to satisfy $a_1 \le \frac{\epsilon}{2\alpha R}$, and $a_2, \ldots, a_{m-1}$ to satisfy, for $i > 1$, $\frac{a_i - a_{i-1}}{a_i} \le \frac{\epsilon}{B}$, where

$$B = \max\left\{\alpha M(R+M), M(L + 1 + 2R\alpha) + \alpha R^2\right\}.$$

If we choose $0 < c < 1$ and set $a_i = (1-c)^{m-i}$ for $i = 1, \ldots, m$, then these conditions are equivalent to $c \le \epsilon/B$ and

$$(1-c)^{m-1} \le \frac{\epsilon}{2\alpha R} \qquad \Leftrightarrow \qquad 1 - \left(\frac{\epsilon}{2\alpha R}\right)^{1/(m-1)} \le c.$$

Thus, it suffices if

$$\begin{aligned}
&\frac{\epsilon}{B} \ge 1 - \left(\frac{\epsilon}{2\alpha R}\right)^{1/(m-1)} \qquad \Leftarrow \qquad \epsilon \ge \frac{B}{m-1}\ln\frac{2\alpha R}{\epsilon} \\
&\Leftarrow \qquad \epsilon \ge \frac{B}{m-1}\max\left\{1, \ln\frac{2\alpha R m}{B}\right\} \qquad \Leftarrow \qquad \epsilon \ge \frac{B\ln 2m}{m-1},
\end{aligned}$$

using the inequality $1 - x \le \ln(1/x)$, which follows from convexity of $\ln(1/x)$. ■

4. Zero Fréchet derivatives with deep compositions

The following theorem is the main result of this section. It shows that if a composition of near-identity maps has zero Fréchet derivatives of a quadratic criterion with respect to the functions in the composition, then the composition minimizes that criterion. That is, all critical points of this kind are global minimizers; there are no saddle points or suboptimal local minimizers in the near-identity region.

Theorem 4 *Consider a distribution P on $\mathcal{X} \times \mathcal{X}$, and define the criterion*

$$Q(h) = \frac{1}{2}\mathbf{E}_{(X,Y)\sim P} \|h(X) - Y\|_2^2 .$$

Define a conditional expectation $h^*(x) = \mathbf{E}[Y|X = x]$, so that h^* minimizes Q. Consider the function computed by an m-layer network $h = h_m \circ \cdots \circ h_1$, and suppose that, for some $0 < \epsilon < 1$ and all i, h_i is differentiable, $\|h_i\| < \infty$, and $\|h_i - \mathrm{Id}\|_L \leq \epsilon$. Suppose that $\|h - h^*\| < \infty$. Then for all i,

$$\inf_{\Delta \in B_1} D_{h_i} Q(h)(\Delta) \leq \frac{-(1-\epsilon)^{m-1}}{\|h - h^*\|} \left(Q(h) - Q(h^*)\right).$$

Thus, if h is a critical point of Q, that is, for all i, $D_{h_i} Q(h) = 0$, we must have $Q(h) = Q(h^*)$.

The theorem defines the expected quadratic loss under an arbitrary joint distribution, but in particular it could be a discrete distribution that is uniform on a training set.

Notice that if h^* satisfies the properties of Theorem 1, then it can be represented as a composition of h_i with the required properties. If it cannot, then the theorem shows that the near-identity region will not contain critical points. The only property we require of h^* is the boundedness condition $\|h - h^*\| < \infty$. From the definition of the induced norm, this implicitly assumes that $h(0) = h^*(0)$ and that h^* is differentiable at 0. In the context of learning embeddings, it seems reasonable to fix the embedding's value at one input vector, and express its value elsewhere relative to that value.

Notice also that, although the theorem requires differentiability of the h_i, it is only important for various derivatives to be defined. In particular, a network with non-differentiable but Lipschitz activation functions, like a ReLU network, could be approximated to arbitrary accuracy by replacing the ReLU nonlinearity with a differentiable one. The conclusions of the theorem apply to any critical point at a differentiable approximation of the ReLU network.

Lemma 5 *Suppose $\|f - \mathrm{Id}\|_L \leq \alpha < 1$.*

1. $(1-\alpha)\|x - y\| \leq \|f(x) - f(y)\| \leq (1+\alpha)\|x - y\|$.
2. *f is invertible and $\left\|f^{-1} - \mathrm{Id}\right\|_L \leq \alpha/(1-\alpha)$.*
3. *For $F(g) = f \circ g$, $\|DF(g) - \mathrm{Id}\| \leq \alpha$, and hence $\|DF(g) - \mathrm{Id}\|_L \leq \alpha$.*

Proof Part 1: The triangle inequality and the Lipschitz property gives

$$\|x - y\| \leq \|f(x) - f(y)\| + \|f(x) - x - (f(y) - y)\| \leq \|f(x) - f(y)\| + \alpha\|x - y\|.$$

Similarly,

$$\|f(x) - f(y)\| \leq \|x - y\| + \|f(x) - x - (f(y) - y)\| \leq (1+\alpha)\|x - y\|.$$

Part 2: For $\alpha < 1$, the inequality $\|f(x) - f(y)\| \geq (1-\alpha)\|x - y\|$ of Part 1 shows that f is invertible. Together with the Lipschitz property, this also shows that

$$\|x - y - (f(x) - f(y))\| \leq \alpha\|x - y\| \leq \frac{\alpha}{1-\alpha}\|f(x) - f(y)\|,$$

which, since $(f^{-1} - \mathrm{Id})(f(x)) = x - f(x)$, gives $\left\|f^{-1} - \mathrm{Id}\right\|_L \le \alpha/(1-\alpha)$.
Part 3: From the definition of $DF(g)$, $\lim_{\Delta\to 0} \|F(g+\Delta) - F(g) - DF(g)(\Delta)\| / \|\Delta\| = 0$. We can write, for any Δ with $\|\Delta\| < \infty$,

$$\begin{aligned}
\|\Delta - DF(g)(\Delta)\| &= \|\Delta + F(g+\Delta) - F(g+\Delta) + F(g) - F(g) + g - g - DF(g)(\Delta)\| \\
&\le \|F(g+\Delta) - F(g) - DF(g)(\Delta)\| \\
&\quad + \|f\circ(g+\Delta) - (g+\Delta) - (f\circ g - g)\| \\
&= o(\|\Delta\|) + \sup_x \frac{\|f\circ(g+\Delta)(x) - (g+\Delta)(x) - (f\circ g - g)(x)\|}{\|x\|} \\
&= o(\|\Delta\|) + \alpha \sup_x \frac{\|\Delta(x)\|}{\|x\|} = o(\|\Delta\|) + \alpha\|\Delta\|.
\end{aligned}$$

Hence, $\|DF(g) - \mathrm{Id}\| \le \alpha$. Since $(DF(g) - \mathrm{Id})$ is a linear functional, this also shows that it is α-Lipschitz:

$$\|DF(g)(\Delta_1) - \Delta_1 - (DF(g)(\Delta_2) - \Delta_2)\| = \|DF(g)(\Delta_1 - \Delta_2) - (\Delta_1 - \Delta_2)\| \le \alpha\|\Delta_1 - \Delta_2\|.$$

■

Proof (of Theorem 4) From the projection theorem,

$$\begin{aligned}
Q(h) &= \frac{1}{2}\mathbf{E}_{(X,Y)\sim P}\|h(X) - Y\|_2^2 \\
&= \frac{1}{2}\mathbf{E}\|h(X) - h^*(X)\|_2^2 + \frac{1}{2}\mathbf{E}_{(X,Y)\sim P}\|h^*(X) - Y\|_2^2 .
\end{aligned}$$

Fix $1 \le i \le m$. To analyze the effect of changing the function h_i on $Q(h)$ by applying the chain rule for Fréchet derivatives, we trace the effect of changing h_i on h by describing h as the result of the composition of a sequence of functionals, which map functions to functions. In particular, we write $h = H_m \circ \cdots \circ H_{i+1} \circ G_i(h_i)$, where $H_j(g) := h_j \circ g$ for $i < j \le m$, $h_i^j := h_j \circ \cdots \circ h_i$ for $i \le j \le m$, and $G_i(g) = g \circ h_{i-1} \circ \cdots \circ h_1$. Now, using the chain rule for Fréchet derivatives,

$$\begin{aligned}
D_{h_i}Q(h) &= \mathbf{E}\left[(h(X) - h^*(X)) \cdot \mathrm{ev}_X \circ D_{h_i}h\right] \\
&= \mathbf{E}\left[(h(X) - h^*(X)) \cdot \mathrm{ev}_X \circ DH_m(h_i^{m-1}) \circ \cdots \circ DH_{i+1}(h_i^i) \circ DG_i(h_i)\right],
\end{aligned}$$

where ev_x is the evaluation functional, $\mathrm{ev}_x(f) := f(x)$. From the definition of the Fréchet derivative, $DG_i(g)$ always satisfies

$$\begin{aligned}
0 &= \lim_{\Delta\to 0} \frac{\|G_i(g+\Delta) - G_i(g) - DG_i(g)(\Delta)\|}{\|\Delta\|} \\
&= \lim_{\Delta\to 0} \frac{\|(g+\Delta)\circ h_{i-1}\circ\cdots\circ h_1 - g\circ h_{i-1}\circ\cdots\circ h_1 - DG_i(g)(\Delta)\|}{\|\Delta\|} \\
&= \lim_{\Delta\to 0} \frac{\|\Delta\circ h_{i-1}\circ\cdots\circ h_1 - DG_i(g)(\Delta)\|}{\|\Delta\|}. \qquad (6)
\end{aligned}$$

The definition of the Fréchet derivative also implies that $DG_i(g)$ is linear, as is the functional F defined by $F(\Delta) = \Delta \circ h_{i-1} \circ \cdots \circ h_1$. If F and $DG_i(g)$ were unequal, progressively scaling down an input on which they differ would scale down the difference by the same amount, contradicting (6). Thus $F = DG_i(g)$, which in turn implies

$$
\begin{aligned}
D_{h_i}Q(h)(\Delta) &= \mathbf{E}\left[(h(X) - h^*(X)) \cdot \mathrm{ev}_X \circ DH_m(h_i^{m-1}) \circ \cdots \circ DH_{i+1}(h_i^i) \circ \Delta \circ h_{i-1} \circ \cdots \circ h_1\right] \\
&= \mathbf{E}\left[(h(X) - h^*(X)) \cdot DH_m(h_i^{m-1}) \circ \cdots \circ DH_{i+1}(h_i^i) \circ \Delta \circ h_{i-1} \circ \cdots \circ h_1(X)\right].
\end{aligned}
$$

For all j, since $(h_j - \mathrm{Id})$ is ϵ-Lipschitz, Lemma 5 implies h_j is invertible. The lemma also implies that, for all j, $\left(DH_j(h_i^{j-1}) - \mathrm{Id}\right)$ is ϵ-Lipschitz, and hence that $DH_j(h_i^{j-1})$ is also invertible. Because these inverses exist, we can define

$$
\Delta = c\left(DH_m(h_i^{m-1}) \circ \cdots \circ DH_{i+1}(h_i^i)\right)^{-1} \circ (h^* - h) \circ (h_{i-1} \circ \cdots \circ h_1)^{-1},
$$

where we pick the scalar $c > 0$ so that $\|\Delta\| = 1$. This choice ensures that

$$
DH_m(h_i^{m-1}) \circ \cdots \circ DH_{i+1}(h_i^i) \circ \Delta \circ h_{i-1} \circ \cdots \circ h_1 = c(h^* - h), \tag{7}
$$

and hence $D_{h_i}Q(h)(\Delta) = -c\mathbf{E}\|h(X) - h^*(X)\|_2^2$. Since $\|\Delta\| = 1$, for all $\gamma > 0$ there is a y with $\|\Delta(y)\| \geq (1-\gamma)\|y\|$. Define $x = (h_{i-1} \circ \cdots \circ h_1)^{-1}(y)$. Then, using the definition of the induced norm and Equation (7), we have

$$
c\|h - h^*\| \geq c\frac{\|h(x) - h^*(x)\|}{\|x\|} = \frac{1}{\|x\|}\left\|DH_m(h_i^{m-1}) \circ \cdots \circ DH_{i+1}(h_i^i) \circ \Delta(y)\right\|.
$$

Recalling that all $\left(DH_j(h_i^j) - \mathrm{Id}\right)$ and $(h_j - \mathrm{Id})$ are ϵ-Lipschitz, we can apply Lemma 5:

$$
\begin{aligned}
c\|h - h^*\| &\geq (1-\epsilon)^{m-i}\frac{\|\Delta(y)\|}{\|x\|} \\
&\geq (1-\epsilon)^{m-i}(1-\gamma)\frac{\|y\|}{\|x\|} \\
&= (1-\epsilon)^{m-i}(1-\gamma)\frac{\|h_{i-1} \circ \cdots \circ h_1(x)\|}{\|x\|} \\
&\geq (1-\epsilon)^{m-1}(1-\gamma).
\end{aligned}
$$

Taking the limit as $\gamma \to 0$ implies the result. ■

5. Bad critical points for sigmoid residual nets

Theorem 4 may be paraphrased to say that residual nets cannot have any bad critical points in the near-identity region, when we consider Fréchet derivatives. In this section, we show that when we consider gradients with respect to the parameters of a fixed-size residual network with sigmoid activation functions, the corresponding statement is not true.

For a depth m, width d and size k, the (m, d, k) $\tanh$ *residual network* N with parameters $\theta = (A_1, ..., A_m, B_1, ..., B_m)$ computes the function $h_\theta \stackrel{\text{def}}{=} h_m \circ ... \circ h_1$, where each layer h_i is defined by $h_i(x) = A_i \tanh(B_i x) + x$, with $A_1, ..., A_m \in \mathbb{R}^{d \times k}$ and $B_1, ..., B_m \in \mathbb{R}^{k \times d}$, and we define $\tanh$ of a vector as the component-wise application of $\tanh$.

To gain an intuitive understanding of the existence of suboptimal critical points, consider the following two properties of networks with $\tanh$ nonlinearities. First, there are finitely many simple transformations (such as permutations of hidden units, or negation of the input and output parameters of a unit) that leave the network function unchanged. Second, apart from these transformations, two networks with different parameter values compute different functions. (This was shown for generic parameter values and $\tanh$ networks of arbitrary depth by Fefferman (1994), and improved by Albertini and Sontag (1992) for the special case of two-layer networks.) Then for any globally optimal parameter value, there is a simple transformation that is also globally optimal. Consider a path between these two parameter values that minimizes the maximum value of the criterion along the path. (It is not hard to construct a scenario in which such a minimax path exists.) The maximizer must be a suboptimal critical point. The proof we give of the following theorem is more direct, relying on specific properties of the $\tanh$ parameterization, but we should expect a similar result to apply to networks with other nonlinearities and parameterizations, provided functions have multiple isolated distinct representations as in the case of $\tanh$ networks.

The proof leverages the fact that, while Theorem 4 rules out the possibility of bad critical points arising from interactions between the layers $h_1, ..., h_m$, they may still arise due to the dynamics of training an individual h_i.

Theorem 6 *For any $\epsilon > 0$, any dimension d, width k and a depth m, for all (m, d, k) tanh residual networks N^* that do not compute the identity function, there is an $R > 0$, and a joint distribution P, over $B_R(\mathbb{R}^d) \times B_R(\mathbb{R}^d)$, such that N^* has parameter θ^* that minimizes $Q(\theta) = \frac{1}{2}\mathbf{E}_{(X,Y)\sim P} \|h_\theta(X) - Y\|_2^2$, and the layers h_i^* of N^* satisfy $\|h_i^* - \mathrm{Id}\|_L \leq \epsilon$, and there is a (m,d,k) tanh residual network N such that N has parameter θ that is a critical point for Q, N has layers h_i that satisfy $\|h_i - \mathrm{Id}\|_L \leq \|h_i^* - \mathrm{Id}\|_L$, but $Q(\theta) > Q(\theta^*)$.*

Proof Let $A_1, ..., A_m, B_1, ..., B_m$ be the parameters of N^* and define f_i by $f_i(x) = A_i \tanh(B_i x) + x$, and $f_{N^*} = f_m \circ ... \circ f_1$.

Let P be any joint distribution over examples (x, y) such that

- $\mathbf{E}(x) = 0$,
- $y = f_{N^*}(x)$ with probability 1, and
- $\mathbf{E}(||y - x||^2) > 0$.

Let N be the network with all-zero parameters. We claim that N a saddle point of Q. Choose a weight w in N, between nodes u and v. Let net be presquashed linear combination of the nodes providing an input to v, so that $v = \tanh(\mathrm{net})$. For a particular (x, y), if we define $Q_{(x,y)}(N) = (f_N(x) - y)^2$, then

$$\frac{\partial Q}{\partial w} = \mathbf{E}_{(x,y)\sim P}\left(\frac{\partial Q_{(x,y)}}{\partial w}\right).$$

Furthermore,

$$\frac{\partial Q_{(x,y)}}{\partial w} = \frac{\partial Q_{(x,y)}}{\partial v}\frac{\partial v}{\partial \mathrm{net}}\frac{\partial \mathrm{net}}{\partial w}.$$

If all of the weights are zero, however, $\frac{\partial Q_{(x,y)}}{\partial v}$ and $\frac{\partial v}{\partial \text{net}}$ do not depend on the input, so that $\frac{\partial Q_{(x,y)}}{\partial w}$ is proportional to $\frac{\partial \text{net}}{\partial w} = u$. However, again, since all of the weights are zero, by induction, there is a component x of the input such that $u = x$, which implies $\mathbf{E}(u) = 0$. Since $\frac{\partial Q_{(x,y)}}{\partial v}$ and $\frac{\partial v}{\partial \text{net}}$ are constant, and $\mathbf{E}\left(\frac{\partial \text{net}}{\partial w}\right) = 0$, we have

$$\frac{\partial Q}{\partial w} = \mathbf{E}_{(x,y)\sim P}\left(\frac{\partial Q_{(x,y)}}{\partial w}\right) = 0,$$

and therefore N is a critical point.

It remains to show that $Q(N) > Q(N^*)$. Since all of the weights of N are 0, N computes the identity function. Since $\mathbf{E}(||y - x||^2) > 0$, this means $Q(N) > 0 = Q(N^*)$. ■

6. Discussion

Informally, Theorem 4 says that, if near-identity behavior on each layer is maintained, for example through regularization or early stopping, optimization with deep residual nets using gradient descent cannot get stuck due to interactions between the layers.

Theorem 4 also has consequences for algorithms that optimize residual networks with a countably infinite number of parameters on each layer, for instance by sequentially adding units. Such algorithms have been studied for two-layer networks; see (Bach, 2014; Bengio et al., 2006; Lee et al., 1996). Indeed, standard denseness results (see the remarks after Theorem 1) show that any downhill direction in function space can be approximated by a single layer network, such as a linear combination of sigmoid units, or of ReLU units. (Notice that, because of the Lipschitz constraint, the class of functions that such a network can compute has bounded statistical complexity.) Thus, for residual networks for which the nonlinear components are computed by standard architectures with a *variable* size, the main result shows that every critical point (with this parameterization) satisfying the Lipschitz constraint is a global optimum.

Typically, however, the size of the hidden layers is fixed. As we have seen in Section 5, while Theorem 4 rules out the possibility of bad critical points due to interaction between the layers, there *can* be bad critical points due to dynamics within an individual layer. On the other hand, a regularizer that promotes in each layer a small Lipschitz norm deviation from the identity function allows the use of a large number of hidden units while avoiding overfitting. This large number of hidden units provides a variety of directions for improvement, and it has been observed that training explores only a small subset of the functions that could be represented with the parameters in each layer (cf., (Denil et al., 2013)). Thus, we might gain useful insight into methods that optimize large networks using a parametric gradient approach by viewing these methods as an approximation to nonparametric, Fréchet gradient methods.

Acknowledgments

Thanks to Dan Asimov for helpful discussions. We gratefully acknowledge the support of the NSF through grant IIS-1619362 and of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

Appendix A. Proof of Lemma 3

Applying the gradient theorem for line integrals to each component of h yields

$$h(y) - h(x) = \int_0^1 Dh(x + t(y-x))(y-x)\,dt,$$

and this, together with smoothness, implies the first inequality:

$$\begin{aligned}
&\|h(y) - (h(x) + Dh(x)(y-x))\| \\
&= \left\| \int_0^1 Dh(x + t(y-x))(y-x)\,dt - Dh(x)(y-x) \right\| \\
&= \left\| \int_0^1 \left(Dh(x + t(y-x)) - Dh(x)\right)(y-x)\,dt \right\| \\
&\le \int_0^1 \left\| \left(Dh(x + t(y-x)) - Dh(x)\right)(y-x) \right\|\,dt \\
&\le \alpha \int_0^1 t\|y-x\|^2\,dt \\
&= \frac{\alpha}{2}\|y-x\|^2.
\end{aligned}$$

For the second, we write

$$\begin{aligned}
h(y) - h(x) &= \int_0^1 Dh(x + t(y-x))(y-x)\,dt \\
&= \int_0^1 Dh(0)(y-x)\,dt + \int_0^1 \left(Dh(0) - Dh(x + t(y-x))\right)(y-x)\,dt,
\end{aligned}$$

hence,

$$\begin{aligned}
\|h(y) - h(x)\| &\le \left\| \int_0^1 Dh(0)(y-x)\,dt \right\| + \alpha \int_0^1 \|x + t(y-x)\|\|y-x\|\,dt \\
&= \|y-x\| + \alpha\|y-x\|R \\
&= (1+\alpha R)\|y-x\|.
\end{aligned}$$

References

Francesca Albertini and Eduardo D Sontag. For neural networks, function determines form. In *Decision and Control, 1992., Proceedings of the 31st IEEE Conference on*, pages 26–31. IEEE, 1992.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *arXiv preprint arXiv:1412.8690*, 2014.

Peter L Bartlett, David P Helmbold, and Philip M Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. *arXiv preprint arXiv:1802.06093*, 2018.

Yoshua Bengio, Nicolas L Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Advances in neural information processing systems*, pages 123–130, 2006.

Misha Denil, Babak Shakibi, Laurent Dinh, Nando de Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pages 2148–2156, 2013.

Charles Fefferman. Reconstructing a neural net from its output. *Revista Matematica Iberoamericana*, 10(3):507–555, 1994.

Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. *NIPS*, 2017.

M. Hardt and T. Ma. Identity matters in deep learning. *ICLR*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

K. Hornik. Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, 4: 251–257, 1991.

Wee Sun Lee, Peter L Bartlett, and Robert C Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122, 2015.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. Learning functions: when is deep better than shallow. *arXiv preprint arXiv:1603.00988*, 2016.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015. Also in COLT'16.