Author Response to Reviews on "Efficient Approximation of Deep ReLU Networks"

We appreciate all reviewers' valuable comments. Here are our response to the major questions raised by the reviewers.

To Reviewer #1:

Q: Line 174, regarding the approximation of f by polynomials and construction of ϕ.

A: We construct ϕ_i's as linear transformations (Line 222) so it can be realized by a simple linear network. To approximate f, we first decompose $f = \sum_{i=1}^{C_{\mathcal{M}}} f_i$ as in Line 248. Our Taylor approximation sub-network approximates each f_i in two components: the first one realizes the linear projection ϕ_i by a linear network and the second one approximates $f_i \circ \phi_i^{-1}$ in the neighborhood U_i by a ReLU network (Theorem 3). We will clarify the realization of ϕ_i's in the next version.

To Reviewer #2:

Q: Step 3: there is an ambiguity in determining the chart to which a point belongs to, how this is solved?

A: We allow a point $\mathbf{x}$ to belong to multiple charts, and the chart determination sub-network determines all the proper charts that $\mathbf{x}$ belongs to (Line 231). Specifically, the U_i's form an open cover of $\mathcal{M}$ (Line 216). Thus, a given input $\mathbf{x}$ can belong to multiple U_i's. For the approximation of f, we associate each U_i with a weight $\rho_i(\mathbf{x})$ from a partition of unity satisfying $\sum_i \rho_i(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{M}$. Then our Taylor approximation sub-network approximates $f_i(\mathbf{x}) = \rho_i(\mathbf{x}) f(\mathbf{x})$ (Line 248). Consequently, the sum of all the outputs from the pairing sub-network (products of the indicator function of U_i and the corresponding Taylor approximation for $f_i(\mathbf{x})$, Line 280) approximates $f(\mathbf{x}) = \sum_{i=1}^{C_{\mathcal{M}}} f_i(\mathbf{x})$.

Q: Taylor approximation has local error guarantees in general, in contrast to the L_∞ approximation used in this paper.

A: While Taylor approximation yields a local error guarantee in each U_i, our L_∞ error bound holds uniformly for $\mathbf{x} \in \mathcal{M}$. A uniform upper bound of all local errors gives rise to the L_∞ error bound (Theorem 4). In this paper, we uniformly bound all local errors and therefore the result is given in the L_∞ error.

Q: Information theoretic bounds - can the authors elaborate? Improve section 4, Ideally.

A: We show our obtained network size matches the lower bound up to log factors (Lines 191 - 193). We will rephrase "information-theoretic bound" as "the lower bound in Theorem 2". We will elaborate on this part in the revision.

We will also improve the technical Section 4 by including more high-level ideas and some graphical illustrations.

To Reviewer #3:

Q: The authors show no experimental results; instead they reference other networks (e.g., VGG, Alexnet, etc.).

A: There have been empirical evidences (VGG, Alexnet, etc.) revealing a huge gap between the network size used in practice and the one predicted by existing theories (Line 52). Therefore, we believe it is not necessary to provide our own experimental results. Our theoretical results bridge this gap by taking low dimensional data structures into consideration, and establish efficient approximation theories for ReLU networks.

Q: Line 14, you say you implement a sub-network but there are no experimental results.

A: "Implementation" here means "analytical construction", which is a standard notion in approximation theory literature. We will use "construct" in the next version to avoid confusion.

Q: Line 48, it is not intuitive what it scales to, please consider rewriting it.

A: We will rephrase it as "To achieve an ϵ uniform approximation error, the number of neurons scales as $\epsilon^{-256\times 256\times 3}$ ([Barron, 1993, Universal approximation bounds]). Setting $\epsilon = 0.1$ gives rise to $10^{256\times 256\times 3}$ neurons." in the revision.

Q: Line 145 (definition 6), to cite https://arxiv.org/pdf/1705.04565.pdf, and comments on the reach.

A: We will add more citations in the next version. Our definition on the reach (Definition 6) is consistent with that in [arXiv:1705.04565]: The set $\mathcal{C}(\mathcal{M})$ (Line 145) contains all points having two closest points in $\mathcal{M}$. Reach is defined as the minimum distance between $\mathcal{M}$ and $\mathcal{C}(\mathcal{M})$. The reach of $\mathcal{M} = \{(x,x) : x \in \mathbb{R}^+\} \cup \{(0,x) : x \in \mathbb{R}^+\}$ is 0, however, this is not a smooth manifold. Our paper considers Riemannian manifold (Line 122) and is therefore smooth. It is generally true that a smooth manifold with a small reach needs a large number of charts (open balls in Line 216).

Q: Line 159 the constant B, and Line 160 the assumption on reach.

A: We will rephrase Line 159 to "There exists $B > 0$ such that, for any $\mathbf{x} \in \mathcal{M}$, we have $|x_i| \leq B$ for $i = 1, \ldots, D$." Assumption 2 says that $\mathcal{M}$ has a positive reach. We will rephrase Assumption 2 to "The reach of $\mathcal{M}$ is $\tau > 0$."

Q: Line 177, it is not clear at all if it is possible to partition a manifold with open sets.

A: We assume the manifold $\mathcal{M}$ is compact (Assumption 1, Line 158). Hence, the existence of a finite open cover is guaranteed by Heine-Borel theorem (See Wikipedia and Folland, Real Analysis, 1999).

Q: Line 272, is each term in the Taylor expansion a different layer of the network?

A: As shown in the proof of Theorem 3 (Lines 499 - 504), each term $\widetilde{f}_{\mathbf{m},\mathbf{n}}$ in the Taylor expansion is approximated by a ReLU network of depth at most $c_1 \log \frac{1}{\delta}$. There are totally $d^n(N+1)^d$ terms in the Taylor expansion (N is chosen as in Line 505). Therefore, the number of terms in the Taylor expansion essentially indicates the width of our Taylor approximation sub-network.