General Bounds for 1-Layer ReLU approximation

Bolton BAILEY & Matus TELGARSKY
Univ. Urbana-Champaign
Urbana IL, USA
{boltonb2,mjt}@illinois.edu

***Abstract*—The popularity of the ReLU has given rise to many neural networks which are piecewise affine. In this work, we show how a refined bound on the number of affine pieces in a single ReLU layer can be used to lower bound the approximation error of a ReLU network. We also demonstrate a method based on Rademacher complexity and random sampling to give an upper bound on the error of optimal approximations for these layers.**

I. INTRODUCTION

With the advent of the age of neural networks, much thought has gone into understanding their power as function approximators. The power of one-nonlinear-layer neural networks serves as a good point of comparison against deep networks, since the former can be seen as the fundamental building-block of the latter. This comparison has caught the attention of researchers who have succeeded in demonstrating functions which are hard to represent with one nonlinear layer, but much easier to represent with two nonlinear layers [1] [2] [3]. These papers share a common thread in that they establish bounds on the number of nodes needed for the approximation of specific functions by one-nonlinear-layer networks. In this paper, we provide bounds that apply not just for single functions, but which have wide applicability. We show an upper bound of the form $\varepsilon \approx 1/n^2$ for a broad class of functions, which we obtain through a multidimensional analogue of a previously known affine piece counting argument. Conversely, we show a lower bound of the form $\varepsilon \approx 1/\sqrt{n}$ based on a probabalistic sampling method.

II. SETTING

We consider functions on the unit ball in d dimensions, approximated under the uniform metric: For functions f, we seek approximations $\tilde{f}$ that minimize $||f - \tilde{f}|| = \sup_{|x| \leq 1} |f(x) - \tilde{f}(x)|$ (This differs from previous work [1] [2] [3], which use L_2 spaces). Here $\tilde{f}$ is a neural network with one hidden layer. That is, the function $\tilde{f}$ has the form

$$\tilde{f}(x) = \sum_{i=1}^{n} ReLU(\langle v_i, x\rangle + b_i)$$

where $v_i \in \mathbb{R}^d, b_i \in \mathbb{R}$. The number of nodes in the hidden layer of the network is denoted by n.

In this paper, we will consider the regime where d is fixed, and we optimize n and the uniform error. We will use f_S to denote the restriction of f to a subset S of its domain. We will use $m(S)$ to denote the Lebesgue measure of set S.

III. AFFINE PIECE COUNTING LOWER BOUND

Our lower bound theorem comes from a result of the authors which describes the number of affine pieces in a network based on the dimensionality of the input space [4, Lemma 2.1]. A univariate version of this lemma appeared in [6], whereafter it was combined with strong convexity assumption to prove a lower bound for deep networks [5]. Here, we use the multivariate generalization to strengthen the convexity-based lower bound proof in the case of shallow networks.

Theorem III.1. *Let f be c-strongly convex on a convex subset S of its domain. Then a one-nonlinear-layer network $\tilde{f}$ with n nodes satisfies $||f - \tilde{f}|| \geq O(1/n^2)$.*

Proof. We have (from Lemma 2.1 of [4]) that the domain of a 1-nonlinear-layer network $\tilde{f}$ can be partitioned into $N_A \leq (e\frac{n}{d} + e)^d$ convex pieces on which $\tilde{f}$ is affine. We therefore have a subset $S' \subseteq S$ of measure $m(S') \geq \frac{m(S)}{N_A} \geq \frac{m(S)}{(e\frac{n}{d}+e)^d}$ contained in a single piece of this partition. Thus, $f_{S'}$ is c-strongly convex, and $\tilde{f}_{S'}$ is affine, and from these facts we will establish $||f - \tilde{f}|| \geq ||f_{S'} - \tilde{f}_{S'}|| = \varepsilon$. Denote $g = f_{S'} - \tilde{f}_{S'}$. Since g is the difference of a c-strongly convex function and an affine function, g is c-strongly convex. Furthermore, g must lie between $[-\varepsilon, \varepsilon]$ on S'. These last two facts tell us that S' is contained within a ball of radius $\sqrt{\frac{4\varepsilon}{c}}$ centered on $\arg\min g$, since at this radius from the minimum of g, the value of g has increased by at least 2ε. This ball has volume $V_d\sqrt{\frac{4\varepsilon}{c}}^d$, where V_d is the d-dimensional circle constant, so:

$$V_d\sqrt{\frac{4\varepsilon}{c}}^d \geq \frac{m(S)}{(e\frac{n}{d} + e)^d}$$

$$\sqrt{\frac{4\varepsilon}{c}} \geq \left(\frac{m(S)}{V_d}\right)^{1/d} \frac{1}{(e\frac{n}{d} + e)}$$

$$\varepsilon \geq \frac{c}{4e}\left(\frac{m(S)}{V_d}\right)^{2/d}\left(\frac{d}{n+d}\right)^2 = O(1/n^2).$$

□

Note that the constant depends on the convexity parameter c, the size of the convex domain $m(S)$, and the dimension d (which we are considering to be fixed).

This theorem can be applied to study a wide array of functions. Any function with a positive definite Hessian at some point x must be strongly convex in a neighborhood of

x. Thus, any smooth function with a local minimum must have a $O(1/n^2)$ approximation rate lower bound. Some examples:

Corollary III.2. *Let $f : x \mapsto e^{-|x|^2}$ on the unit sphere. A one-nonlinear-layer network $\tilde{f}$ with n nodes satisfies $||f - \tilde{f}|| \geq O(1/n^2)$.*

Proof. Since $-f$ is strongly convex near 0, Theorem IV.1 can be applied. □

Corollary III.3. *Let $x \mapsto |x|_2$ on the unit sphere. A one-nonlinear-layer network $\tilde{f}$ with n nodes satisfies $||f - \tilde{f}|| \geq O(1/n^2)$.*

Proof. While f is not strongly convex, we can restrict f to the hyperplane $H = \{|x| < 1 : x_1 = \frac{1}{2}\}$. This restricted function f_H in $d - 1$ dimensions is smooth and has a minimum at $(\frac{1}{2}, 0, \ldots, 0)$, so we can apply the theorem in this subspace. □

IV. Upper Bound

Our upper bound technique comes from an application of Rademacher complexity. For a function f, the technique first involves determining a measure μ_f on the set of parameterizations for a single ReLU $\{v \in R^d, b \in R\}$ such that

$$f(x) = k \int ReLU(\langle x, v \rangle + b) d\mu_f(v, b)$$

for a constant k. For example, in the case of the 2-norm function $f(x) = |x|_2 = \sqrt{\sum_i x_i^2}$, a measure $\mu_f = \sigma$ is suitable, where σ is a distribution where v is uniform over the unit sphere and $b = 0$. In general, for radial functions of the form $f(x) = g(|x|)$, where g is a polynomial, the density function of μ_f can also be chosen to be radial, and is also a polynomial. Thus, it is possible to produce μ_f for a wide array of radial functions through approximation of Taylor series.

We simplify a bit and assume μ_f is a distribution (we can use a Jordan decomposition of μ_f to make this rigorous): over a selection of n draws $(v_1, b_1), \ldots, (v_n, b_n)$ from μ_f, the function

$$\hat{f} = \frac{k}{n} \sum_{i=1}^{n} ReLU(\langle x, v_i \rangle + b_i),$$

which is representable as an n-node shallow network, should approximate f as $n \to \infty$. We use a Rademacher complexity argument to reason that

$$||\hat{f} - f|| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{n}}\right)$$

with probability $1 - \delta$, where the constant depends on k. We then instantiate this bound with a fixed δ and argue by the probabalistic method that the optimal n-node approximation to f satisfies

$$||\hat{f} - f|| \leq O\left(\sqrt{\frac{1}{n}}\right).$$

V. Conclusions and Future Work

In this paper, we have established lower and upper bounds that can be applied broadly to the approximation of functions by shallow networks. The lower bound and upper bounds provided in the previous two sections are $O(\frac{1}{n^2})$ and $O(\frac{1}{\sqrt{n}})$ respectively. Noting the gap between these asymptotics is a power of 4, future work in this direction should include tightening these asymptotic bounds to match each other. Furthermore, we could expand our view to consider how the dimension d impacts the constants in these bounds.

References

[1] I. Safran & O. Shamir *Depth Separation in ReLU Networks for Approximating Smooth Non-Linear Functions.* CoRR (2016).

[2] R. Eldan & O. Shamir *The Power of Depth for Feedforward Neural Networks.* CoRR (2015).

[3] A. Daniely *Depth Separation for Neural Networks.* CoRR (2017).

[4] B. Bailey & M. Telgarsky *Size-Noise Tradeoffs in Generative Networks.* Advances in Neural Information Processing Systems (2018), 6490-6500.

[5] S. Liang & R. Srikant, *Why deep neural networks for function approximation?.* CoRR, (2016).

[6] M. Telgarsky, *Benefits of depth in neural networks.* Journal of Machine Learning Research **49** (2016), 1-23.