On the shapes of functions generated by random neural networks

David Holmes

December 6, 2022

Abstract

We consider functions from $\mathbb{R} \to \mathbb{R}$ output by a neural network with 1 hidden activation layer, arbitrary width, and ReLU activation function. We assume that the parameters of the neural network are chosen uniformly at random with respect to various probability distributions, and compute the expected distribution of the points of non-linearity. We use these results to explain why the network may be biased towards outputting functions with simpler geometry, and why certain functions with low information-theoretic complexity are nonetheless hard for a neural network to approximate.

Contents

1 Introduction

It has been suggested [VPCL18, MSVP$^+$19, MVPSL20] that neural networks are biased in favour of outputting 'simple' functions. The above papers interpret simplicity in an information-theoretic fashion, suggesting that functions output by neural networks tend to have small information-theoretic complexity. The converse is not true in general; for example, it is relatively hard for a neural network to learn a periodic function such as the integral of $x \mapsto (-1)^{\lfloor x/n \rfloor}$ compared to its low information-theoretic complexity.

We study some more geometric properties of functions which help to predict how easily they are generated by a neural network. We consider functions defined on an interval $(-R, R) \subseteq \mathbb{R}$, and write $\mathsf{PL}_r(\mathbb{R})$ for the space of piecewise-affine-linear (PL) real-valued functions on $(-R, R)$. Given positive integer w we write $\mathsf{PL}_{R,\leq w}(\mathbb{R})$ for the subspace of those PL functions with at most w points of non-linearity.

We consider a neural network with one hidden activation layer, and ReLU activation function. Then the space of functions output by the neural network is exactly the space $\mathsf{PL}_{\leq w}$. As such, one way of 'randomly' generating elements of $\mathsf{PL}_{\leq w}$ is to 'randomly' choose the weights of the neural network. Our main result, approximately stated, has two flavours:

1. For R fairly small, if a function in $\mathsf{PL}_{\leq w}$ is generated by a random neural network with ReLU activation, then it is likely to have much fewer than w points of non-linearity.

2. For R large, the points of non-linearity behave as if they are sampled from a probability distribution with mass function of shape

$$x \mapsto \min(\frac{1}{4}, \frac{1}{4x^2}). \tag{1.0.1}$$

In particular, whether r is large or small, a periodic zigzag function such as the integral of $x \mapsto (-1)^{\lfloor x/n \rfloor}$ is hard to approximate relative to its information-theoretic complexity.

Of course, the exact statements depend on how we choose the parameters of our neural network. Perhaps surprisingly, we will see that the distribution of the points of non-linearity depends very heavily on the distribution of the parameters. If the bias towards simple functions underlies generalisation properties of over-parameterised neural networks (as proposed in [MVPSL20], this may help to explain why some gradient descent schemes generalise better than others (as they approximate random sampling with respect to distributions more heavily favouring simple functions).

1.1 Statement of main results

Our neural network has parameters taking values in some measurable subset of Θ of a real vector space. The probabilities of seeing a given number of points of non-linearity turn out to be highly dependent on the *shape* of the parameter space Θ, but independent of the *size* of Θ. We will consider two different 'shapes', one 'rectangular' and the other 'spherical'.

1.1.1 Rectangular parameter space

We consider neural networks with one hidden activation layer, defining functions from $(-R, R) \subseteq \mathbb{R}$ to $\mathbb{R}$. We write w for the width (a positive integer). Our parameter space Θ is naturally a product of 4 pieces:

1. a linear part in the first layer; this gives a subspace of $\mathbb{R}^w$, denoted Θ_L
2. a translation part in the first layer; this again gives a subspace of $\mathbb{R}^w$, denoted Θ_T
3. a linear part in the final layer; this again gives a subspace of $\mathbb{R}^w$, denoted Θ'_L
4. a translation part in the final layer; this gives a subspace of $\mathbb{R}$, denoted Θ'_T.

So $\Theta = \Theta_L \times \Theta_T \times \Theta'_L \times \Theta'_T$. Now it turns out that Θ'_L and Θ'_T have no effect on the number of points of non-linearity (outside some measure-zero subset of Θ'_L, which we ignore). So it suffices to describe the spaces Θ_L and Θ_T, and their probability measures.

For this we fix a positive real number T. We define

$$\Theta_L = \Theta_T = [-T, T]^w, \tag{1.1.1}$$

a box of side-length T and dimension w, centred at $0 \in \mathbb{R}^w$. We equip it with the lebesgue measure. In other words,

- the 'translation' part of the first layer is chosen uniformly at random between $-T$ and T at each neuron;
- the scaling factor at each neuron in the first layer is chosen uniformly at random between $-T$ and T.

It is easy to see (lemma 2.1) that a PL function generated in this way has at most w points of non-linearity. In fact, the number of points of non-linearity follows a binomial distribution:

Proposition 1.1. *Given any $w' \in \{1, 2, \ldots, w\}$, the probability of a function generated by this neural network having exactly w' points of non-linearity is*

$$\binom{w}{w'} \mathcal{P}^{w'} (1 - \mathcal{P})^{w-w'}$$

where $\binom{w}{w'}$ is the binomial coefficient, and

$$\mathcal{P} = \begin{cases} \frac{R}{2} & \textit{if } 0 < R \leq 1 \\ 1 - \frac{1}{2R} & \textit{if } R \geq 1. \end{cases} \tag{1.1.2}$$

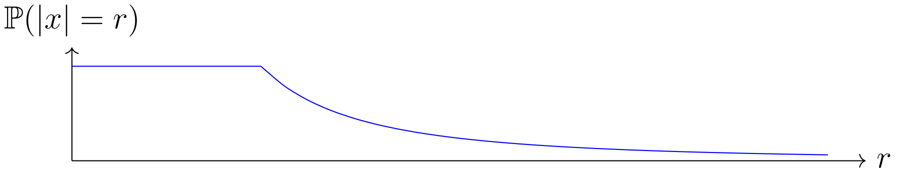
In particular, the expected number of points of non-linearity is given by

$$\mathbb{E}(w') = w\mathcal{P} < w.$$

1.1.2 Functions on an unbounded domain

Suppose that we use the same parameter space Θ, but we now view our PL functions as having domain $\mathbb{R}$. Then for almost all $\theta \in \Theta$, the PL function will have w points of non-linearity. However, the distribution of these points is far from uniform. Differentiating the above cresult, we find that the probability density function of the distribution of these w points is given by

$$\mathbb{P}(|x| = r) = \begin{cases} \frac{w}{2} & \text{if } 0 \leq r \leq 2 \\ \frac{w}{2r^2} & \text{if } 2 \leq r \end{cases} \tag{1.1.3}$$



1.1.3 Spherical parameter space

Our setup here is similar. We again fix a positive real number T, and define

$$\Theta_T = [-T, T]^w, \tag{1.1.4}$$

a box of side-length $2T$ and dimension w, centred at $0 \in \mathbb{R}^w$.

The 'spherical' part comes in Θ_L; we define

$$\Theta_L = \{L \in \mathbb{R}^w : |L| \leq T\}, \tag{1.1.5}$$

a sphere of radius T around the origin in $\mathbb{R}^w$. Again, we equip Θ_L with the Lebesgue measure.

In this context it does not seem so easy to compute the exact probability of a given number of points of non-linearity occurring. However, at least for $R \leq 1$, we can compute the expected number of points of non-linearity.

Proposition 1.2. *Assume $0 < R \leq 1$. For w even we have*

$$\mathbb{E}(w') = \frac{Rw2^w}{(w+1)\pi}\binom{w-1}{w/2}^{-1} \sim R\sqrt{2w/\pi}. \tag{1.1.6}$$

For odd w we have

$$\mathbb{E}(w') = \frac{Rw^2}{2^{w-1}(w+1)}\binom{w-1}{(w-1)/2} \sim R\sqrt{2w/\pi}. \tag{1.1.7}$$

Here $\sim$ means that the ratio tends to 1 as w tends to infinity.

Since $\sqrt{w}$ is much smaller than w, this indicates that such functions tend strongly to having few points of linearity.

1.2 Possible generalisations and extensions

1.3 Training data

If the training data forces w points of non-linearity, then of course they will occur with probability 1 among parameters fitting the training data. However, the fact that we generally work with highly over-parametrised models says that this will in general not be the case. We expect that similar results will hold (and be provable with similar techniques) in the presence of training data, as long as certain 'over-parametrisation' conditions are satisfied.

1.4 Higher dimensions

Instead of looking at functions taking values on an interval $(-R, R)$ in the reals, it is natural to look at function on $(-R, R)^i$ for some positive integer i. Here the locus of points where the function is not linear will not be finite; as such, instead of using the cardinality of that set as a measure of simplicity, we will instead use its Hausdorff measure. We believe that similar results can be shown, by similar methods.

1.5 Different activation functions

Our results are valid not just for the ReLU activation function, but in fact any piecewise-linear activation function which has a unique point of non-linearity at the origin. Generalising to other PL activation functions will require no major changes. For differentiable activation functions which are asymptotically linear, we can replace the measure of the locus of points of non-linearity by (for example) the integral of the square of the largest eigenvalue of the Hessian. Again, we expect similar results, but different techniques will be required to prove them.

These results strongly suggest that neural networks (at least with one layer and ReLU activation) will not generalise well when approximating e.g. periodic or polynomial functions. To rectify this, one could assign some neurons a periodic or polynomial activation function. Heuristic computations suggest that the dilation is a reasonable predictor of bias in the presence of polynomial activation functions (i.e. functions with lower dilation are more likely to be chosen).

2 DNNs with ReLU activation; rectangular norm

In this section we fix

$$\varphi\colon \mathbb{R} \to \mathbb{R}; x \mapsto \max(x, 0),$$

though the same analysis and results hold for any continuous activation function which fails to be linear exactly at 0.

Given $\theta \in \Theta$, we write $D_\theta \subseteq (-R, R)$ for the set of points of non-linearity of the function given by the parameters θ.

Lemma 2.1. *$\#D_\theta \leq w$, and this maximum can be achieved.*

Proof. Suppose the image of X is not contained in a coordinate hyperplane of $\mathbb{R}^w$. Then D_θ is exactly the image under a linear map of the intersection of the image of X with the coordinate hyperplanes in $\mathbb{R}^w$, of which there are at most w.

On the other hand, if the image of X is contained in the intersection of exactly w' of the coordinate hyperplanes, then the image of X hits at most $w - w'$ other coordinate hyperplanes. □

Fix $w' \in \{0, 1, \ldots, w\}$. Our first goal is to compute $\mathbb{P}(\#D_\theta = w')$.

A point in $c \in \mathbb{R}^w$ is chosen uniformly at random in a box around 0 of side-length $2T$. Another point $l \in \mathbb{R}^w$ is chosen uniformly at random in a box around 0 of side-length $2RT$. Then we consider the line segment in $\mathbb{R}^w$ joining $c - l$ and $c + l$, and we want to compute the probability of this segment meeting any of the coordinate hyperplanes; given $1 \leq i \leq w$ write $\mathcal{P}_i$ for the probability of our line segment meeting the ith coordinate hyperplanes; this is independent of i, so we also write it $\mathcal{P}$.

Lemma 2.2. *If $R \geq 2$ then $\mathcal{P} = 1 - \frac{1}{2R}$. If $0 < R \leq 1$ then $\mathcal{P} = \frac{r}{R}$.*

Proof. Without loss of generality, $i = 1$. Then for fixed c the probability of intersecting the point $x_1 = 0$ is

$$\max(1 - \frac{|c_1|}{RT}, 0).$$

Integrating over $|c_1|$ from 0 to T yields the result. □

Since the probabilities of hitting the various axes are independent, we deduce

Proposition 2.3. *For $w' \in \{0, 1, \ldots, w\}$ the probability of a function generated by this neural network having exactly w' points of non-linearity is*

$$\mathbb{P}(\#D_\theta = w') = \binom{w}{w'} \mathcal{P}^{w'} (1 - \mathcal{P})^{w-w'}. \tag{2.0.1}$$

The expected number of points of non-linearity is given by

$$\mathbb{E}(\#D_\theta) = \sum_{w' \in \{0,\ldots,w\}} \frac{\mu(\Theta_{w'})}{\mu(\Theta)} w' = w\mathcal{P}. \tag{2.0.2}$$

3 DNNs with ReLU activation; operator norm

Just as in the rectangular case, we have

Lemma 3.1. *$\#D \leq w$, and this maximum can be achieved.*

Fix $w' \in \{0, 1, \ldots, w\}$. Let

$$\Theta_{w'} := \{\theta \in \Theta : \mu(D_\theta) = w'\} \subseteq \Theta. \tag{3.0.1}$$

Note that $\mu(\Theta) = t^N$. To compute the bias exactly is to compute the individual $\mu(\Theta_{w'})$, which seems somewhat tricky. But by a simple application of classical results from geometric probability we will be able to compute the expected measure of D_θ. More precisely, we define

$$\mathbb{E}(w') = \sum_{w' \in \{0,\ldots,w\}} \frac{\mu(\Theta_{w'})}{\mu(\Theta) = t^N} w'. \tag{3.0.2}$$

For example, if $\mathbb{E}(w') \approx w$ this would tell us that most choices of parameters yield $\mu(D) = w$. On the other hand, if $\mathbb{E}(w') \approx 0$ this would tell us that most choice of parameter give an affine-linear function.

Proposition 3.2. *Assume $0 < R \leq 1$. For w even we have*

$$\mathbb{E}(w') = \frac{Rw2^w}{(w+1)\pi}\binom{w-1}{w/2}^{-1} \sim R\sqrt{2w/\pi}. \tag{3.0.3}$$

For odd w we have

$$\mathbb{E}(w') = \frac{Rw^2}{2^{w-1}(w+1)}\binom{w-1}{(w-1)/2} \sim R\sqrt{2w/\pi}. \tag{3.0.4}$$

Here $\sim$ means that the ratio tends to 1 as w tends to infinity.

Proof. As before, only the first component $\theta_1 \colon \mathbb{R} \to \mathbb{R}^w$ of the map $P(\theta)$ has any impact on the number $\mu(D_\theta)$; more precisely, $\mu(D_\theta)$ is the number of intersection points of the image of $[-1, 1]$ under θ_1 with the coordinate hyperplanes in $\mathbb{R}^w$ (excluding the measure-zero case where the image crosses the intersection of two or more coordinate hyperplanes). Since the measure on Θ is a product of the measures onto factors, we may simply ignore the second factor of θ.

We now relate the problem to a variation on Buffon's needle. The image of $[-R, R]$ is a line segment in $\mathbb{R}^w$, with centre a point in $[-T, T]^w$ chosen uniformly at random, and endpoint chosen uniformly at random in a sphere of radius RT around that centre. We want to compute the expected number of intersection points with the coordinate hyperplanes.

For now we fix the length $2s \in [0, 2RT]$ of the needle, and compute the expectation; later we will integrate over s. By additivity of expectations, we are reduced to computing the expected number $\frac{1}{w}\mathbb{E}(w')$ of intersection points with a single coordinate hyperplane. By symmetry, the expected number of intersection points with a coordinate hyperplane is the same as the expected number of intersection points of a needle of length $2s$, dropped uniformly at random in the plane, with the subset

$$\{x \in \mathbb{R}^w : x_1 \in 2T\mathbb{Z}\}. \tag{3.0.5}$$

By [KR97, page 130] this expectation is given by[1]

$$\mathbb{E} = \frac{\Omega_1 \Omega_{w-1}}{w\Omega_w}\frac{s}{T} \tag{3.0.6}$$

where for a non-negative integer k we have

$$\Omega_{2k} = \frac{\pi^k}{k!} \tag{3.0.7}$$

and

$$\Omega_{2k+1} = \frac{2^{2k+1}\pi^k k!}{(2k+1)!}. \tag{3.0.8}$$

[1] We write Ω where Klain and Rota write ω, to make the distinction from the width w clearer.

We find for even w that

$$\mathbb{E} = \frac{2^w s}{w\pi T}\binom{w-1}{w/2}^{-1} \tag{3.0.9}$$

and for odd w that

$$\mathbb{E} = \frac{s}{2^{w-1}T}\binom{w-1}{(w-1)/2}. \tag{3.0.10}$$

To simplify subsequent computations, we write $\mathbb{E}' = T\mathbb{E}/s$, which depends only on w. To complete the computation of the expectations we must integrate over $s \in [0, RT]$. However, we do not integrate with respect to the uniform distribution on $[0, RT]$; rather we want the endpoint of our needle to be chose uniformly in a sphere. As such, the expectation for hitting one hyperplane is

$$\frac{1}{w}\mathbb{E}(w') = B(w, TR)^{-1}\int_{s=0}^{RT} \frac{s}{T}\mathbb{E}' S(w, s) ds \tag{3.0.11}$$

where

$$B(w, TR) = \frac{\pi^{w/2}}{\Gamma(\frac{w}{2}+1)}(TR)^w$$

is the volume of a ball of radius TR and dimension w, and

$$S(w, s) = \frac{2\pi^{w/2}}{\Gamma(\frac{w}{2})}s^{w-1}$$

is the surface area of a ball of dimension w and ratios s. This turns into

$$\begin{aligned}\frac{1}{w}\mathbb{E}(w') &= \frac{\Gamma(\frac{w}{2}+1)}{\pi^{w/2}(TR)^w}\int_{s=0}^{RT} \frac{s}{T}\mathbb{E}'\frac{2\pi^{w/2}}{\Gamma(\frac{w}{2})}s^{w-1} ds \\ &= R\frac{w}{w+1}\mathbb{E}'.\end{aligned} \tag{3.0.12}$$

For the asymptotics we apply the central binomial coefficient formula

$$\binom{2k}{k} \sim \frac{4^k}{\sqrt{k\pi}}, \tag{3.0.13}$$

and for even w we also use

$$\binom{2k-1}{k-1} = \frac{1}{2}\binom{2k}{k}. \tag{3.0.14}$$

□

References

[KR97] Daniel A Klain and Gian-Carlo Rota. *Introduction to geometric probability.* Cambridge University Press, 1997.

[MSVP$^+$19] Chris Mingard, Joar Skalse, Guillermo Valle-Pérez, David Martínez-Rubio, Vladimir Mikulik, and Ard A. Louis. Neural networks are a priori biased towards boolean functions with low entropy, 2019.

[MVPSL20] Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is sgd a bayesian sampler? well, almost. 2020.

[VPCL18] Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions, 2018.

David Holmes
Mathematisch Instituut, Universiteit Leiden, Postbus 9512, 2300 RA Leiden, Netherlands
E-mail address: `holmesdst@math.leidenuniv.nl`