Approximation of Continuous Functions by ReLU Nets

Will Troiani

October 2019

1 Introduction

Feed-forward neural nets with a single hidden layer can approximate continuous functions on a compact domain if the hidden layer is allowed to be arbitrarily wide [5], [6]. Recently though [4, §3] there has been interest in feed-forward networks with bounded *widths* rather than *depth*. The goal of this talk is to discuss the approximation of continuous functions on compact domains by ReLU nets with an upper bound on the widths (Theorem 1 and Remark 1).

2 Feed-forward Neural Nets, ReLU activations, and Max-min strings

Definition 1. *A* ***ReLU net*** $\mathcal{N}$ *consists of*

- *a (finite) sequence of* ***widths*** $d_{in} = d_1, d_2, ..., d_n = d_{out}$, *where each* $d_i \in \mathbb{N}$, d_{in} *is the* ***input width****, and* d_{out} *the* ***output width****,*
- *a sequence of affine functions* $(A_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}})_{i=1}^{n-1}$,

Associated to every ReLU net $\mathcal{N}$ is a function

$$f_{\mathcal{N}} : A_n \mathrm{ReLU}_{d_{n-1}} A_{n-1} ... \mathrm{ReLU}_{d_1} A_1 : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}^{d_{\mathrm{out}}} \tag{1}$$

where

$$\begin{aligned} \mathrm{ReLU}_k &: \mathbb{R}^k \to \mathbb{R}^k \\ (x_1, ..., x_k) &\mapsto \big(\max\{0, x_1\}, ..., \max\{0, x_k\}\big) \end{aligned}$$

The subscript k on ReLU_k will be dropped when its value is clear from context. The proof that ReLU nets can be used to approximate continuous functions with compact domains will not use ReLU nets on the nose, but instead will use functions which can be written as particular compositions involving affine functions as well as the max and min functions:

Definition 2. *A **max-min string** is a function $g : \mathbb{R}^n \to \mathbb{R}^m$ which can be written as*

$$g = \sigma_{L-1}(l_L, \sigma_{L-2}(l_{L-1}, ...(\sigma_2(l_3, \sigma_1(l_2, l_1))...)) \tag{2}$$

where each $l_i : \mathbb{R}^n \to \mathbb{R}^m$ is an affine function, and each σ_i is either the component-wise max *function:*

$$\begin{aligned} \max : \mathbb{R}^{2m} &\to \mathbb{R}^m \\ (x_1, ..., x_m, y_1, ..., y_m) &\mapsto (\max\{x_1, y_1\}, ..., \max\{x_m, y_m) \end{aligned}$$

or the component-wise min *function:*

$$\begin{aligned} \min : \mathbb{R}^{2m} &\to \mathbb{R}^m \\ (x_1, ..., x_m, y_1, ..., y_m) &\mapsto (\min\{x_1, y_1\}, ..., \min\{x_m, y_m\}) \end{aligned}$$

Lemma 1. *Let $g : \mathbb{R}^n \to \mathbb{R}^m$ a max-min string and $K \subseteq \mathbb{R}^n$ a compact set. Then there exists a ReLU net $\mathcal{N}$ such that for all $x \in K$, $f_{\mathcal{N}}(x) = g(x)$.*

Proof. It can be assumed without loss of generality that K lies in the positive orthant, that is,

$$K \subseteq \mathbb{R}^n_+ := \{(x_1, ..., x_n) \in \mathbb{R}^n \mid 1 \le i \le n \Rightarrow x_i \ge 0\}$$

To see this, say that K does *not* lie in the positive orthant. Then let $T : K \to K'$ be a translation where K' does lie in the positive orthant, such a translation exists as K is compact. Then by the theorem (once it has been proved) there exists $\mathcal{N}$ such that $f_{\mathcal{N}} = gT^{-1}$, which implies $f_{\mathcal{N}}T = g$, and $f_{\mathcal{N}}T$ is a ReLU net.

So assume K lies in the positive orthant and write

$$g = \sigma_{L-1}(l_L, \sigma_{L-2}(l_{L-1}, ...(\sigma_2(l_3, \sigma_1(l_2, l_1))...))$$

where $l_i : \mathbb{R}^n \to \mathbb{R}^m$.

The key observation is that if $a, b \in \mathbb{R}$, then $\max\{a, b\}$ can be calculated by adding a to $\max\{0, b - a\}$, with a similar trick used for min.

Define the following affine functions:

$$\begin{aligned} A_1 : \mathbb{R}^n &\to \mathbb{R}^m \\ x &\mapsto (x, l_1(x)) \end{aligned}$$

and if $i = 2, ..., L$:

$$\begin{aligned} A_i : \mathbb{R}^{n+m} &\to \mathbb{R}^{n+m} \\ (x, y) &\mapsto \begin{cases} (x, y - l_i(x)) & \sigma_{i-1} = \max \\ (x, -y + l_i(x)) & \sigma_{i-1} = \min \end{cases} \end{aligned}$$

Notice that for $i = 2, ..., L$, the functions A_i admit inverses given by:

$$\begin{aligned} A_i^{-1} &: \mathbb{R}^{n+m} \to \mathbb{R}^{n+m} \\ (x, y) &\mapsto \begin{cases} (x, y + l_i(x)) & \sigma_{i-1} = \max \\ (x, -y + l_i(x)) & \sigma_{i-1} = \min \end{cases} \end{aligned}$$

Then the function

$$h := \text{ReLU } A_L^{-1} \text{ ReLU } A_L \ ... \ A_1^{-1} \text{ ReLU } A_1$$

is equal to the graph of g. Thus $\pi h = g$, where $\pi : \mathbb{R}^{2m} \to \mathbb{R}^m$ is the map $(x_1, ..., x_n, y_1, ..., y_m) \mapsto (y_1, ..., y_m)$, and πh is a ReLU net. □

3 Approximation of Continuous Functions by ReLU Nets, Finite Case

It follows from Lemma 1 that in order to approximate a continuous function $f : K \to \mathbb{R}^m$ using a ReLU net, it suffices to approximate f with a max-min string. The goal of section 4 below is to show how this can be done. The proof will require some technical analysis, but the broad idea is highlighted well by considering the case when K is a finite set. In fact, in the finite case, f can be computed exactly using a ReLU net:

Lemma 2. *Let $\epsilon \in \mathbb{R}_{>0}$ and $f : S \to \mathbb{R}^m$ a function with S finite. Then there exists a max-min string $g : \mathbb{R}^n \to \mathbb{R}^m$ such that $f = g$.*

The following Lemma will be used:

Lemma 3. *Let $f : S \to \mathbb{R}^m$ be a function with $S \subseteq \mathbb{R}^n$ a finite set. For any $t \in \mathbb{R}_{\geq 0}$ and any $x \in \mathbb{R} \setminus S$ there exists an affine function $l = (l_1, ..., l_m) : \mathbb{R}^n \to \mathbb{R}^m$ such that*

- *$l(x) = (0, ..., 0)$,*
- *$t < l_i(s)$ for all $s \in S$*

Proof. Many such functions l can be defined, but in order to reflect the construction which will be used to prove the case where S is an arbitrary compact set, the case when $n = 2$ and $m = 1$ will be considered and a particular construction will be shown which perhaps is not the most obvious one.

Let $A, B \in \mathbb{R}^2$ be points such that ΔAxB forms a triangle, $S \cap \Delta AxB = \{x\}$ and S lies in the infinite planar sector $\angle AxB$, ie

$$S \subseteq \angle AxB := \{\big(\gamma_1(B - x), \gamma_2(C - x)\big) \mid \gamma_1, \gamma_2 \geq 0\}$$

see figure 1. Then let l be the affine function such that $l(x) = 0$ and $l(A) = l(B) = t$. See figure 2 □

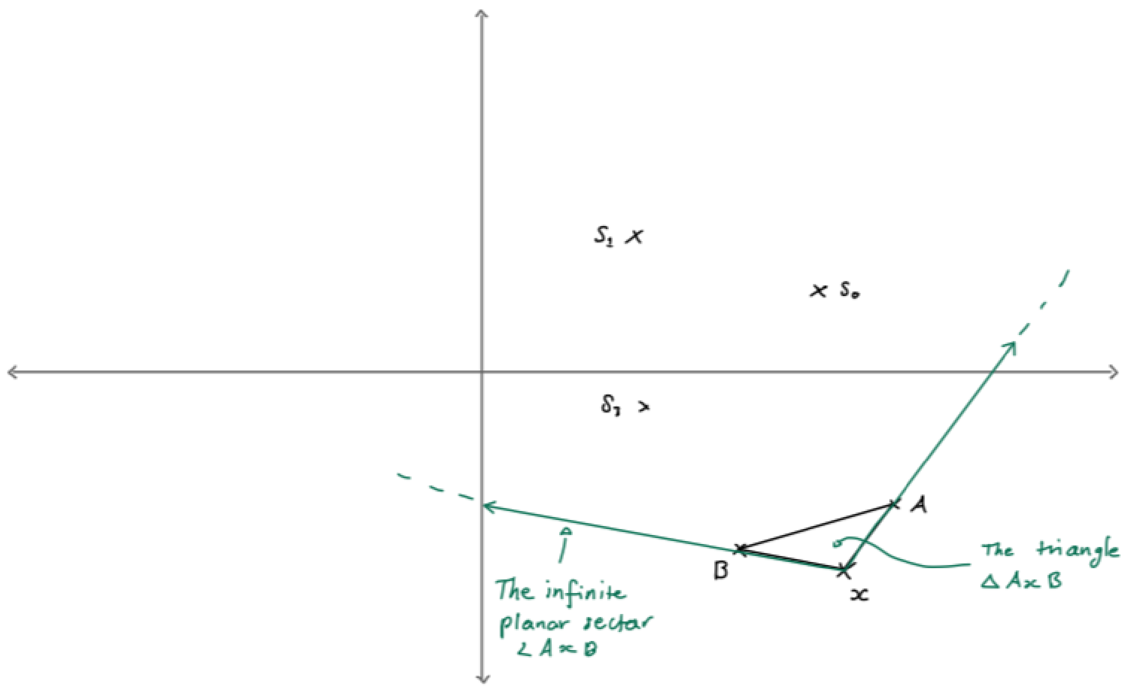

Figure 1: An example of an appropriate choice of triangle for the proof of Lemma ??.

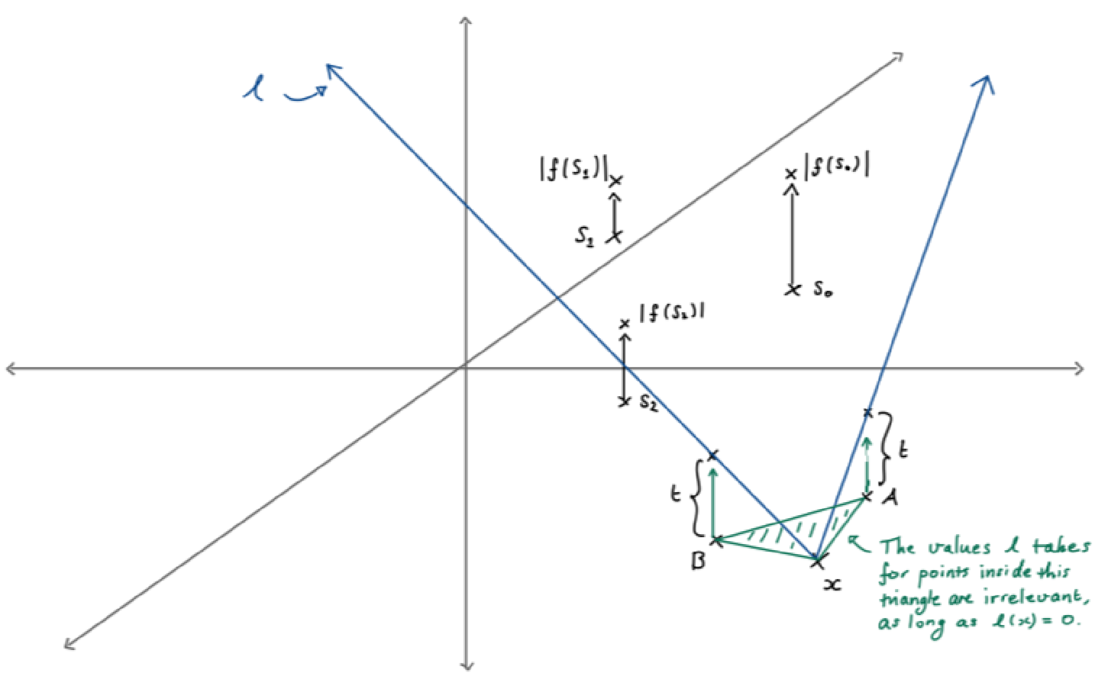

Figure 2: The affine function l.

Proof of Lemma 2. If $S = \{s\}$ contains a single element, then the max-min string $g(s) = f(s)$ can be used. So say S contains at least two elements. Fix an element $s \in S$ and let $g : S \setminus \{s\} \to \mathbb{R}^m$ be a max-min string satisfying $\forall s' \in S \setminus \{s\}$, $g(s') = f(s')$. Set

$$t = \max_{s' \in S \setminus \{s\}} \{|g(s') - f(s)|\}$$

and let $l = (l_1, ..., l_m) : \mathbb{R}^n \to \mathbb{R}^m$ be an affine function such that $l(s) = 0$ and for all $s' \in S \setminus \{s\}$, $l_i(s') > t$, which exists by Lemma 3. Write $f = (f_1, ..., f_m)$, then

$$\forall s' \in S \setminus \{s\}, \forall i = 1, ..., m, \qquad f_i(s) - l_i(s') < g_i(s') < f_i(s) + l_i(s') \tag{3}$$

Define the max-min string:

$$\hat{g} = \max\{\min\{g, f(s) + l\}, f(s) - l\}$$

which is such that $\hat{g}(s') = f(s')$ for all $s' \in S$. Indeed,

$$\begin{aligned} \hat{g}(s) &= \max\{\min\{g(s), f(s) + l(s)\}, f(s) - l(s)\} \\ &= \max\{\min\{g(s), f(s)\}, f(s)\} \\ &= f(s) \end{aligned}$$

and if $s' \in S \setminus \{s\}$,

$$\begin{aligned} \hat{g}(s') &= \max\{\min\{g(s'), f(s) + l(s')\}, f(s) - l(s')\} \\ &= \max\{g(s'), f(s) - l(s')\} \\ &= g(s') = f(s') \end{aligned}$$

where the first and second equality follow from 3. □

4 Approximation of Continuous Functions by ReLU Nets, General Case

The goal of this section is to prove that ReLU nets can approximate continuous functions $f : K \to \mathbb{R}$, where K is compact (for a formal statement see Theorem 1 below). This will be done by approximating a continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ on a compact subset $K \subseteq \mathbb{R}^2$. Again, a ReLU net which approximates f itself will not be directly constructed, but a max-min string will be instead, which is sufficient by Lemma 1.

In the finite case, a max-min string $g : S \setminus \{s\} \to \mathbb{R}^m$ was extended to $\hat{g} : S \to \mathbb{R}^m$. The existence of $\hat{g}$ used crucially that there existed an affine function $l : \mathbb{R}^n \to \mathbb{R}^m$ such that each entry of $l(s')$ was larger than the corresponding entry of $g(s')$, for all $s' \in S \setminus \{s\}$, and was such that $l(s) = 0$. The key realisation in the case where K is an arbitrary compact set is the existence of an analogous affine function l:

Definition 3. *The* ***inverse modulus of continuity of*** ϵ *of a continuous function* $f : \mathbb{R}^2 \to \mathbb{R}$ *in the domain* $K \subseteq \mathbb{R}^2$ *is*

$$w_{f,K}^{-1}(\epsilon) = \sup\{\delta \in \mathbb{R} \mid \forall x, y \in K,\ |x - y| \leq \delta \Rightarrow |f(x) - f(y)| \leq \epsilon\}$$

Lemma 4. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be continuous,* $\epsilon > 0$*,* $r > r' > 0$*. Let* $S_r := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = r\}$ *be the circle of radius* r*. Let* $X, Y \in S_r$ *be such that* $|X - Y| \leq r$ *and* $X', Y' \in S_{r'}$ *be such that* X', X, Y, Y' *are collinear, and let* $Z \in S_{r'}$ *be the point on the arc connecting* X' *to* Y' *and the straight line which passes through the origin and the midpoint of the straight line connecting* X *and* Y*. See Figure 3. Let* L *denote the*

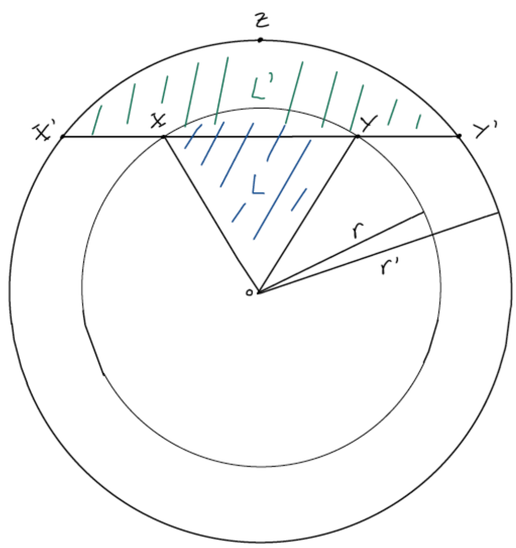

Figure 3: The configuration described in Lemma 4

minor sector of S_r *induced by the radii* $0X$ *and* $0Y$ *and let* L' *denote the minor segment of* $S_{r'}$ *induced by the chord which connects* X' *to* Y'*. If*

$$\mathrm{Diam}(L') \leq w_{f,B_{r'}(0)}^{-1}(\epsilon)$$

then there exists an affine function $l : \mathbb{R}^2 \to \mathbb{R}$ *such that*

- $l(x) \leq \epsilon$ *for all* $x \in L'$*,*
- $|f(x) - f(Z)| \leq l(x) + \epsilon$ *for all* $x \in L$*.*

Proof. Let l be the affine function such that $l(Z) = 0$ and $l(X') = l(Y') = \epsilon$. Let $x \in L$. Denote by $Z' \in \mathbb{R}^2$ the point which is collinear to X', Y', X, Y and is also collinear to Z, x. Then as l is affine and $|x - Z| \leq |Z' - Z|$, it follows that $l(x) \leq l(Z') = \epsilon$.

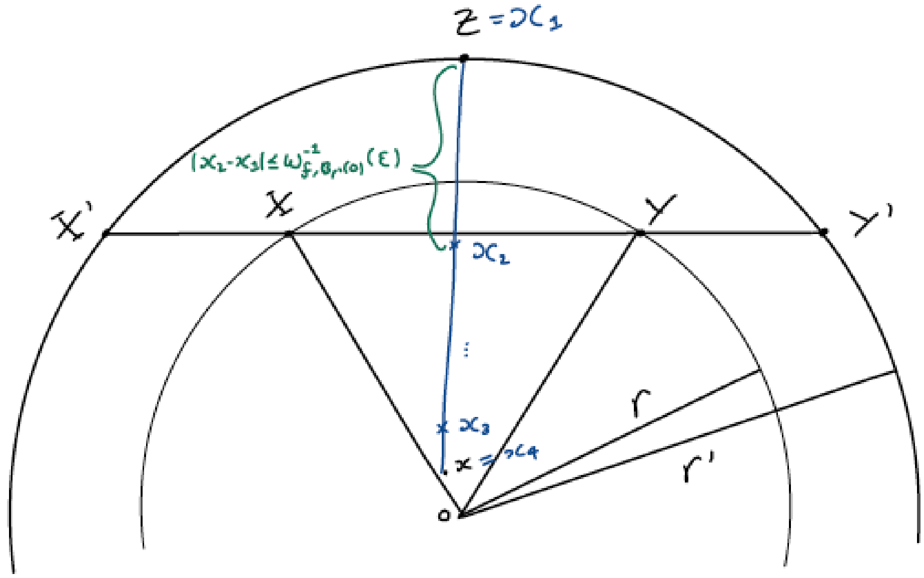

Figure 4: The points $x_1, ..., x_n$ in the proof of Lemma 4

Say $x \in L'$. Then write $|x - Z| = kw^{-1}_{f,B_{r'}(0)}(\epsilon) + \rho$ where $k \in \mathbb{N}$ and $0 \leq \rho < w^{-1}_{f,B_{r'}(0)}(\epsilon)$. Then

$$\begin{aligned} |x - Z| &= kw^{-1}_{f,B_{r'}(0)}(\epsilon) + \rho \\ \Rightarrow |x - Z| &\geq kw^{-1}_{f,B_{r'}(0)}(\epsilon) \end{aligned}$$

and so,

$$\frac{\epsilon|Z - x|}{w^{-1}_{f,B_{r'}(0)}(\epsilon)} \geq k\epsilon \tag{4}$$

Also, consider a sequence of points $Z = x_1, x_2, ..., x_{k+1} = x$ such that $|x_{i+1} - x_i| = w^{-1}_{f,B_{r'}(0)}(\epsilon)$ for $i = 1, ..., k-1$ and $|x_{k+1} - x_k| < w^{-1}_{f,B_{r'}(0)}(\epsilon)$. See Figure 4. Then

$$\begin{aligned} |f(x) - f(Z)| &= |f(x_1) - f(x_2) + f(x_2) - f(x_3) + ... + f(x_2) - f(x_1)| \\ &\leq |f(x_1) - f(x_2)| + ... + |f(x_2) - f(x_1)| \\ &< k\epsilon + \epsilon \end{aligned}$$

and so,

$$|f(x) - f(Z)| < k\epsilon + \epsilon \tag{5}$$

Equations 4 and 5 together imply

$$|f(x) - f(Z)| < \frac{\epsilon|x - Z|}{w^{-1}_{f,B_{r'}(0)}(\epsilon)} + \epsilon$$

Thus it remains to show:

$$\frac{\epsilon|x-Z|}{w^{-1}_{f,B_{r'}(0)}(\epsilon)} \leq l(x)$$

This comes down to the fact that the slope of l in the direction of the vector $x - Z$ can be bounded below by $\frac{\epsilon}{w^{-1}_{f,B_{r'}(0)}(\epsilon)}$. More precisely, consider the following parametrisation of the straight line which intersects Z and x:

$$\begin{aligned} c : \mathbb{R} &\to \mathbb{R}^2 \\ s &\mapsto Z + \frac{s}{|x-Z|}(x-Z) \end{aligned}$$

Then the function $lc : \mathbb{R} \to \mathbb{R}$ is affine, and in fact is linear as

$$lc(0) = l(Z) = 0$$

So for all $y \in \mathbb{R}$, $lc(y) = my$ for some gradient m. Let $W \in \mathbb{R}^2$ be the point which intersects the line parametrised by c and the line which passes through X and Y. Then since $|W - Z| \leq w^{-1}_{f,B_{r'}(0)}(\epsilon)$ and $l(W) = \epsilon$, it follows that:

$$m = \frac{l(W) - l(Z)}{|W-Z|} = \frac{\epsilon}{|W-Z|} \geq \frac{\epsilon}{w^{-1}_{f,B_{r'}(0)}(\epsilon)}$$

and so:

$$\frac{\epsilon|x-Z|}{w^{-1}_{f,B_{r'}(0)}(\epsilon)} \leq m|x-Z| = lc(|x-Z|) = l(x)$$

which completes the proof. □

Corollary 1. *In the setting of Lemma 4, but with every instance of ϵ replaced by $\frac{\epsilon}{2}$, if there exists a max-min string $g : \mathbb{R}^2 \to \mathbb{R}$ which approximates f on L' then there exists a max-min string $\hat{g} : \mathbb{R}^2 \to \mathbb{R}$ which approximates f on $L \cup L'$.*

Proof. Let $l : \mathbb{R}^2 \to \mathbb{R}$ be an affine function such that

- $l(x) \leq \frac{\epsilon}{2}$, for all $x \in L'$,
- $|f(x) - f(Z)| \leq l(x) + \frac{\epsilon}{2}$, for all $x \in L$

the existence of which is guaranteed by Lemma 7. Then define the max-min string

$$\hat{g} = \max\{\min\{g, f(Z) + l\}, f(Z) - l\}$$

On L, we have $f \leq g + \epsilon$ and

$$f \leq f(Z) + l + \epsilon$$

so

$$\begin{aligned} f &\le \min\{g+\epsilon, f(Z)+l+\epsilon\} \\ &= \min\{g, f(Z)+l\}+\epsilon \\ &\le \max\Big\{\min\{g, f(Z)+l\}, f(Z)-l\Big\}+\epsilon \\ &= \hat{g}+\epsilon \end{aligned}$$

Similarly, since $g-\epsilon \le f$ and

$$f(Z)-l-\epsilon \le f$$

it follows that

$$\begin{aligned} \hat{g}-\epsilon &= \max\Big\{\min\{g, f(Z)+l-\epsilon\}, f(Z)-l\Big\}-\epsilon \\ &= \max\{\min\{g-\epsilon, f(Z)+l-\epsilon\}, f(Z)-l-\epsilon\} \\ &\le \max\{g-\epsilon, f(Z)-l-\epsilon\} \\ &\le f \end{aligned}$$

On L', by construction:

$$f(Z)-\frac{\epsilon}{2} \le f(Z)-l \le \hat{g} \le f(Z)+l \le f(Z)+\frac{\epsilon}{2}$$

and for all $x \in L'$

$$|f(x)-f(Z)| \le \frac{\epsilon}{2}$$

by definition of $w^{-1}_{f,B_{r'}(0)}(\frac{\epsilon}{2})$. Thus:

$$\begin{aligned} |\hat{g}(x)-f(x)| &= |\hat{g}(x)-f(Z)+f(Z)-f(x)| \\ &\le |\hat{g}(x)-f(Z)|+|f(Z)-f(x)| \\ &\le \frac{\epsilon}{2}+\frac{\epsilon}{2} \\ &= \epsilon \end{aligned}$$

□

Next we turn to two geometric Lemmas:

Lemma 5. *Consider two concentric circles, S_R and $S_{R'}$ with $R' > R$. Consider a chord intersecting points X and Y on S_R and extend this chord so that it intercepts $S_{R'}$, at X' and Y' say (see Figure 5). Then*

$$|X'Y'|^2 = 4(R'^2 - R^2) + |XY|^2$$

Proof. Let P denote the midpoint of the line XY. Then $|P| = \sqrt{R^2 - \frac{1}{4}|XY|^2}$ (again, see Figure 5). Thus

$$\left(\sqrt{R^2-\frac{1}{4}|XY|^2}\right)^2 + \frac{1}{4}|X'Y'|^2 = R'^2$$

from which, the result follows. □

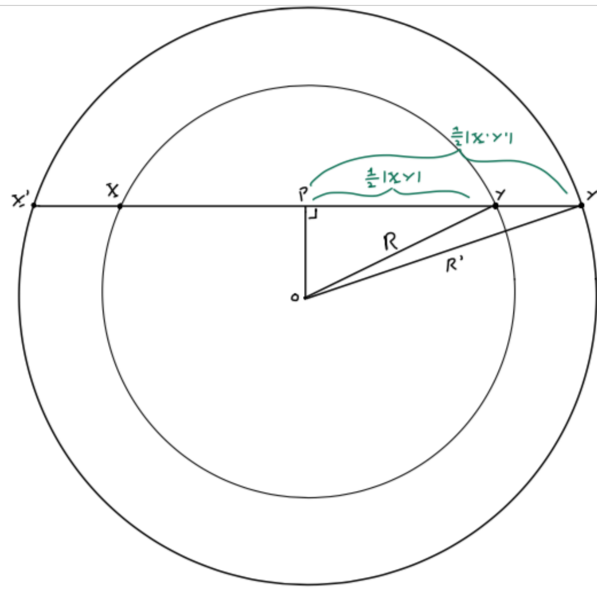

Figure 5: The geometry of Lemma 5

Lemma 6. *Let X, Y be points on S_R such that $0 < |X - Y| \leq R$. Let τ_X and τ_Y be straight lines tangent to S_R respectively at X and Y. Let T denote the intersection of τ_X and τ_Y. Then* $\mathrm{Diam}(\Delta XTY) = |X - Y|$

Proof. If $|X - Y| = R$ then the triangle XOY is equilateral. It follows from this that angle $\angle TXY$ and $\angle TYX$ are both equal to $30°$. As $|X - Y|$ decreases, both angles $\angle TXY$ and $\angle TYX$ decrease. So if $|X - Y| \leq R$ the angle $\angle XTY$ is obtuse, which implies that $\mathrm{Diam}(\Delta XTY) = |X - Y|$. See Figure 6. $\square$

Lemma 7. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be continuous, $\epsilon > 0$, $R > 0$. Assume $w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})$ and let $R' > w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})$. Assume further that there exists a max-min string $g : \mathbb{R}^2 \to \mathbb{R}$ which approximates f on $B_{R'}(0)$. Let R'' be*

$$R'' := \sqrt{R'^2 + \frac{w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})}{\sqrt{2}R'}}$$

If $R'' < R$ then there exists a max-min string $\hat{g} : \mathbb{R}^2 \to \mathbb{R}$ which approximates f on $B_{R''}(0)$, and if $R'' \geq R$ then there exists a max-min string $\hat{g} : \mathbb{R}^2 \to \mathbb{R}$ which approximates f on $B_R(0)$.

Proof. To ease notation let $a_{R'} := \frac{w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})}{\sqrt{2}R'}$. Let $X, Y \in \mathbb{R}^2$ lie on the circle of radius R' which satisfy

$$|XY| = w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})(1 - a_{R'})^{\frac{1}{2}}$$

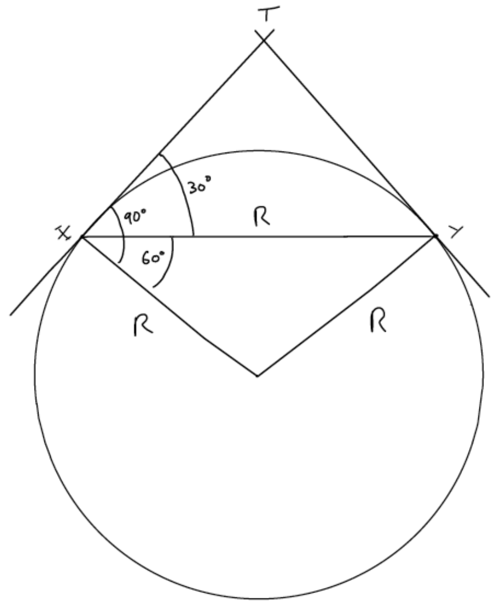

Figure 6: The geometry of Lemma 6

such a pair (X, Y) exists as

$$|XY|^2 = w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})^2(1 - a^2_{R'}) < w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2}) \leq R'$$

We consider first the case when $R'' < R$. Let $X', Y' \in \mathbb{R}^2$ lie on the circle of radius R'' which intersect the line segment connecting X and Y. Let L denote the minor segment on the circle of radius R' induced by the chord which connects X' to Y' and let L' denote the minor sector on the circle of radius R' induced by the radii $0X$ and $0Y$. Let Z denote the intersection of the circle of radius R'' and the line which connects 0 and the midpoint of the line XY. See Figure 7. By Lemma 5 and a direct calculation,

$$|X'Y'| = \sqrt{4(R''^2 - R'^2) + |XY|^2} = w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})$$

Moreover, $w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2}) < R' < R''$, so by Lemma 6:

$$\text{Diam}(L) \leq w^{-1}_{f,B_R(0)}(\frac{\epsilon}{2})$$

Thus the hypotheses of Corollary 1 are satisfied.

The case when $R'' \geq R$ is almost identical but the pair (X', Y') are taken to intersect the ball of radius R instead of the ball of radius R''. $\square$

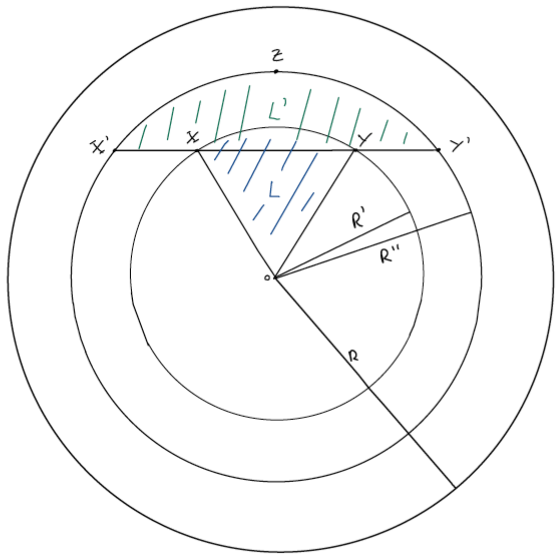

Figure 7: Configuration of the balls in Lemma 7

This brings us to the main Theorem:

Theorem 1. *Let $\epsilon \in \mathbb{R}_{>0}$ and $f : \mathbb{R}^2 \to \mathbb{R}$ be continuous, and $K \subseteq \mathbb{R}^2$ compact. Then there exists a max-min string $g : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$\forall x \in K,\ |g(x) - f(x)| \leq \epsilon$$

Proof. Since compact subsets of $\mathbb{R}^2$ are bounded, it suffices to consider the case when $K = B_R := \{x \in \mathbb{R}^2 \mid |x| \leq R\}$, for some $R \in \mathbb{R}$. If $R \leq w^{-1}_{f,B_R(0)}\left(\frac{\epsilon}{2}\right)$ then the max-min string $g : \mathbb{R}^2 \to \mathbb{R}$ which is the constant function $g \equiv f(0)$ suffices. Suppose $w^{-1}_{f,B_R(0)}\left(\frac{\epsilon}{2}\right) < R$ and $f : B_R(0) \to \mathbb{R}$ are given. Consider the following set

$$\mathscr{X} := \{R' \in \mathbb{R} \mid \text{there exists a max-min string } g \text{ which approximates } f \text{ on } B_R(0)\}$$

We claim that $R \in \mathscr{X}$, recall the notation from Lemma 7 that $a_R := \frac{w_f^{-1}\left(\frac{\epsilon}{2}\right)^2}{\sqrt{2}R}$. Suppose for a contradiction that $R \notin \mathscr{X}$, then $\mathscr{X}$ is bounded above. Let $s^2 = \sup \mathscr{X}^2$ and $R' \in \mathscr{X}$ to be such that

$$s^2 - a_s < R'^2$$

Notice that $s \notin \mathscr{X}$ as if $s \in \mathscr{X}$ then Lemma 7 can be used to find a strictly greater value in $\mathscr{X}$. As $R' < s$, it follows that $a_s < a_{R'}$ and

$$s^2 < R'^2 + a_s < R'^2 + a_{R'}$$

thus

$$s < \sqrt{R'^2 + a_{R'}}$$

but by Lemma 7, $\sqrt{R'^2 + a_{R'}} \in \mathscr{X}$, contradicting that s is a supremum. $\square$

Remark 1. *As mentioned earlier, Theorem 1 generalises to continuous functions* $f : K \to \mathbb{R}^m$ *where* $K \subseteq \mathbb{R}^n$. *Ie, given* $\epsilon > 0$ *there exists a max-min string* $g : \mathbb{R}^n \to \mathbb{R}^m$ *such that for all* $x \in K$,

$$|f(x) - g(x)| < \epsilon$$

Lemma 1 then implies that there exists a ReLU net with input width n *and all other widths equal to* $n + m$ *which approximates* f. *This gives the promised upper bound on the widths. In fact a lower bound can be proved as well, see [3, §3].*

References

[1] K. Weierstrass, *Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen*,Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, 1885.

[2] M. H. Stone, *The Generalised Weierstrass Approximation Theorem*, Mathematics Magazine, 21 (4): 167–184, 1948.

[3] B. Hanin, M. Sellke, *Approximating Continuous Functions by ReLU Nets of Minimal Width*, arXiv:1710.11278, 2018

[4] B. Hanin, *Universl function approximation by deep neural nets with bounded width and ReLU acivations.* arXiv preprint arXiv: 1708.02691, 2017

[5] G. Cybenko, *Approximation by superpositions of a sigmoidal function.* Mathematics of Control, Signals, and Systems (MCSS), 2(4):303-314, 1989

[6] K. Hornik, M. Stinchcombe, H. White, *Multilayer Feedforward Networks are Universal Approximators.* Neural networks, 2(5): 359 - 366, 1989.