THE APPROXIMATION OF CONTINUOUS FUNCTIONS BY MULTILAYER PERCEPTRONS

IAN GLOVER

1. Preliminaries

Definition 1.1. A *perceptron* is a unit consisting of a finite number of inputs and a single output. The output is a function of the inputs.

A *linear perceptron* is a perceptron which given inputs $x_i,\ i \in \{1, \dots, n\}$ has output $\mathcal{O}$ given by

$$\mathcal{O} = \sum_{i \in I} w_i x_i + \theta$$

where $\theta \in \mathbb{R}$ is the *bias* and the $w_i \in \mathbb{R}$ are the weights. ie, the output is a linear combination of the inputs.

A *sigmoid perceptron* is a perceptron which given inputs $x_i,\ i \in \{1, \dots, n\}$ has output $\mathcal{O}$ given by

$$\mathcal{O} = \frac{1}{1 + \exp\left(-\sum_{i \in I} w_i x_i + \theta\right)}$$

where $\theta \in \mathbb{R}$ is the *bias* and the $w_i \in \mathbb{R}$ are the weights.

Definition 1.2. For a continous function $f : [a, b] \to \mathbb{R}$ we define the metric $\|f\|_{[a,b]}$ by

$$\|f\|_{[a,b]} = \int_{[a,b]} |f(x)|\, dx$$

The result we aim to prove is that given a bounded continuous function from a bounded subset of the real line we can approximate it to arbitary accuracy using a multilayer perceptron network with one hidden layer. This will follow immediately from

Theorem. *Given a continuous function $f : [a, b] \to \mathbb{R}$ and $\varepsilon > 0$ n 1-input perceptrons with output functions $\mathcal{O}_{1i} : \mathbb{R} \to \mathbb{R},\ i \in \{1, \dots, n\}$ and one n-input perceptron with output function $\mathcal{O}_{21} : \mathbb{R}^n \to \mathbb{R}$ and such that*

$$\|f - \mathcal{O}_{21}(\mathcal{O}_{11}, \dots, \mathcal{O}_{1n})\|_{[a,b]} < \varepsilon$$

To do this we will show firstly that any continuous function from a closed bounded subset of the real line can be approximated by a linear combination of indicator functions. Secondly we show that when restricted to a finite subset of the real line we can approximate an indicator function by a pair of sigmoid perceptrons. Putting these together will allow us to prove the above result.

Date: 24 November 1999.

2. The approximation of the continuous functions

In this section we aim to show that any continuous function from a bounded subset of the real line can be approximated by a linear combination of indicator functions, with respect to the $\|.\|_{[a,b]}$ metric given above.

Lemma 2.1. *If $f : [a, b] \to \mathbb{R}$ is a continuous function then f is uniformly continuous.*

Lemma 2.2. *If $f : [a, b] \to \mathbb{R}$ is a continuous function then the image f is closed and bounded.*

These results are standard results from any first course in analysis and will not be proved here.

Proposition 2.3. *Any continuous function $f : [0, 1] \to \mathbb{R}$ can be approximated by a finite combination of indicator functions.*

Proof. By lemma 2.1 f is uniformly continuous. Therefore given $\varepsilon > 0$ we can find $\delta(\varepsilon)$ such that $|f(x) - f(y)| < \varepsilon$ for all x, y such that $|x - y| < \delta(\varepsilon)$. Choose $N \in \mathbb{N}$ such that $N > \frac{1}{\delta}$, and let $S_k = [\frac{k}{N}, \frac{k+1}{N}]$ for $k \in \{0, \ldots, N-1\}$. Define $g_{f,\varepsilon}$ by

$$g_{f,\varepsilon} = \sum_{k=0}^{N-1} f\left(\frac{2k+1}{2N}\right) \mathcal{I}_S(x)$$

Now if $x \in \mathcal{I}_S(x)$ then $\left|x - \frac{2k+1}{2N}\right| < \delta(\varepsilon)$ and so

$$\left|f(x) - f\left(\frac{2k+1}{2N}\right)\right| < \varepsilon$$

Thus

$$\begin{aligned}
\|f - g_{f,\varepsilon}\|_{[0,1]} &= \int_{[0,1]} |f(x) - g_{f,\varepsilon}(x)|\, dx \\
&= \sum_{k=0}^{N-1} \int_{S_k} \left|f(x) - f\left(\frac{2k+1}{2N}\right)\right| dx \\
&< \sum_{k=0}^{N-1} \int_{S_k} \varepsilon dx \\
&= \sum_{k=0}^{N-1} \varepsilon \frac{1}{N} \\
&= \varepsilon
\end{aligned}$$

□

Corollary 2.4. *Any continuous function $f : [a, b] \to \mathbb{R}$ can be approximated by a finite combination of indicator functions.*

Proof. Apply proposition 2.3 to $f\left(\frac{x-a}{b-a}\right)$. □

3. Approximation of Indicator Functions by Perceptrons

The next step is to show that we can approximate a step function using the linear combination of a pair of sigmoid perceptrons.

Recall from definition 1.1 that a sigmoid perceptron with one input x has output function:

$$\mathcal{O}_{(\alpha,\beta)}(x) = \frac{1}{1+e^{-\alpha(x+\beta)}}$$

Now consider the heaviside function $H(x)$ given by

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

Lemma 3.1. *Given $\varepsilon > 0$ we can find $\alpha(\varepsilon)$ such that*

$$\left\| H - \mathcal{O}_{(\alpha,0)} \right\|_{[a,b]} < \varepsilon$$

Proof. Consider the condition $\mathcal{O}_{(\alpha,0)}(x) < \varepsilon$ for all $x < -k$ for some $k > 0$ to satisfy this we get:

$$\begin{aligned} \varepsilon &> \mathcal{O}_{(\alpha,0)}(x) \text{ for all } x < -k, \text{ some } k \in \mathbb{R}^+ \\ \frac{1+\varepsilon}{\varepsilon} &< e^{-\alpha x} \text{ for all } x < -k, \text{ some } k \in \mathbb{R}^+ \\ &< e^{-\alpha k} \\ \text{So } \alpha k &\leq \ln\varepsilon - \ln(1+\varepsilon) \end{aligned}$$

This can be achieved by setting

$$\alpha = \frac{1}{k}[\ln\varepsilon - \ln(1+\varepsilon)]$$

It is clear that if we do this then, by symmetry:

$$\mathcal{O}_{(\alpha,0)}(x) > 1 - \varepsilon \text{ for all } x > k$$

If we restrict H to $[a,b]$ where there is $k>0$ such that $[-k,k]\subset[a,b]$ we now get

$$\begin{aligned}\left\|H-\mathcal{O}_{(\alpha,0)}\right\|_{[a,b]} &= \int_a^b |H(x)-\mathcal{O}_{\alpha,0}|\,dx\\ &= \int_a^{-k} |0-\mathcal{O}_{\alpha,0}|\,dx\\ &\quad+\int_{-k}^{k} |H(x)-\mathcal{O}_{\alpha,0}|\,dx\\ &\quad+\int_{k}^{b} |1-\mathcal{O}_{\alpha,0}|\,dx\\ &= \int_a^{-k} \varepsilon dx\\ &\quad+\int_{-k}^{k} |H(x)-\mathcal{O}_{\alpha,0}|\,dx\\ &\quad+\int_{k}^{b} \varepsilon dx\\ &< (-k-a)\varepsilon+\int_{-k}^{k} dx+(b-k)\varepsilon\\ &= (b-a-2k)\varepsilon+2k\end{aligned}$$

If we set $k=\varepsilon$ we get $\left\|H-\mathcal{O}_{(\alpha,0)}\right\|_{[a,b]}<(b-a)\varepsilon$. ☐

Now consider $[c,d]\subset[a,b]$ then $\mathcal{I}_{[c,d]}(x)=H(x-a)-H(x-b)$ so we get the following inmediate corollary to the above

Corollary 3.2. *If $[c,d]\subset[a,b]$ then given $\varepsilon>0$ there exists α such that*

$$\left\|\mathcal{I}_{[c,d]}-(\mathcal{O}_{(\alpha,-a)}-\mathcal{O}_{(\alpha,-b)})\right\|_{[a,b]}<\varepsilon$$

4. The approximation of continuous functions by MLPs

We are now in a position to prove the theorem in stated section 1, namely

Theorem 4.1. *Given a continuous function $f:[a,b]\to\mathbb{R}$ and $\varepsilon>0$ n 1-input perceptrons with output functions $\mathcal{O}_{1i}:\mathbb{R}\to\mathbb{R},\ i\in\{1,\dots,n\}$ and one n-input perceptron with output function $\mathcal{O}_{21}:\mathbb{R}^n\to\mathbb{R}$ and such that*

$$\|f-\mathcal{O}_{21}(\mathcal{O}_{11},\dots,\mathcal{O}_{1n})\|_{[a,b]}<\varepsilon$$

Proof. By proposition 2.3 there exists a function

$$g_{f,\varepsilon}=\sum_{k=0}^{N-1} f\left(\frac{2k+1}{2N}\right)\mathcal{I}_{[\frac{k}{N},\frac{k+1}{N}]}(x)$$

such that $\|f-gf,\varepsilon\|_{[a,b]}<\varepsilon$.

By corollary 3.2 we have can find α such that

$$\left\|\mathcal{I}_{[\frac{k}{N},\frac{k+1}{N}]}-(\mathcal{O}_{(\alpha,-\frac{k}{N})}-\mathcal{O}_{(\alpha,-\frac{k+1}{N})})\right\|_{[a,b]}<\frac{\varepsilon}{N}\text{ for all }k\in\{0,\dots,N-1\}.$$

Thus we define $P(x) : [a,b] \to \mathbb{R}$ by

$$P(x) = \sum_{k=0}^{N-1} f\left(\frac{2k+1}{2N}\right)\left(\mathcal{O}_{(\alpha,-\frac{k}{N})}(x) - \mathcal{O}_{(\alpha,-\frac{k+1}{N})}(x)\right)$$

Then

$$\begin{aligned}
\|f-P\|_{[a,b]} &\leq \|f-g_{f,\varepsilon}\|_{[a,b]} + \|g_{f,\varepsilon}-P\|_{[a,b]} \\
&< \varepsilon + \sum_{k=0}^{N-1} f\left(\frac{2k+1}{2N}\right)\left\|\mathcal{I}_{[\frac{k}{N},\frac{k+1}{N}]} - \left(\mathcal{O}_{(\alpha,-\frac{k}{N})}(x) - \mathcal{O}_{(\alpha,-\frac{k+1}{N})}(x)\right)\right\|_{[a,b]} \\
&< \varepsilon + \frac{\varepsilon}{N}\sum_{k=0}^{N-1} f\left(\frac{2k+1}{2N}\right) \\
&< \varepsilon\left(1 + \frac{N}{N}\sup_{x\in[a,b]} f(x)\right) \\
&< \varepsilon\left(1 + \sup_{x\in[a,b]} f(x)\right)
\end{aligned}$$

By lemma 2.2 $\left|\sup_{x\in[a,b]} f(x)\right| < K$ for some $K \in \mathbb{R}$.

Thus to construct the network to approximated f we have a hidden layer with $2N$ single input perceptrons with weight α and biases $-\frac{k}{N\alpha}$ for $k \in \{1,\ldots,N\}$ (two for each k). We then have single output perceptron with no bias and weights as given by the definition of $P(x)$. □

From this we get the following

Corollary 4.2. *Any piecewise continuous function from a closed bounded subset of the real line can be approximated to arbitrary accuracy by a multilayer perceptron network with one hidden layer.*

Corollary 4.3. *Any piecewise continuous function from a closed bounded subspace of* $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *can be approximated to arbitrary accuracy by a multilayer perceptron network with one hidden layer.*

Corollary 4.4 (Reasonably but not 100% sure of this)**.** *Any Riemann integrable function from* $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *can be approximated to arbitrary accuracy by a multilayer perceptron network with one hidden layer.*

E-mail address: inglover@iname.com