1 Theorem 1

We state the theorem and proof from [**?**] as follows;

Let $1 \leq d \leq s, n \geq 1$ be integers, $f : \mathbb{R}^s \to \mathbb{R}$ and $\phi : \mathbb{R}^d \to \mathbb{R}$. A *generalized translation network* with n neurons evaluates a function of the form $\sum_{k=1}^{n} a_k \phi(A_k(\cdot)+\mathbf{b}_k)$ where the *weights* A_ks are $d \times s$ real matrices, the *thresholds* $\mathbf{b}_k \in \mathbb{R}^d$ and the *coefficients* $a_k \in \mathbb{R}$ $(1 \leq k \leq n)$. The set of all such functions (with a fixed n) will be denoted by $\Pi_{\phi;n,s}$.

We attempt to approximate f by elements of $\Pi_{\phi;n,s}$ on the domain $[-1,1]^s$. For the $d = 1$ case we have $\Pi_{\phi;n,s}$ simply the collection of shallow networks with n neurons in their hidden layer.

This construction will not provide the best approximation, but will instead provide the optimal order of approximation. We have also, that the weights A_k's and the thresholds $\mathbf{b}_k$'s will be determined independently of the target function f.

We observe a notable assumption that we are able to sample from the target function at all prescribed points without noise.

1.1 Statement

We can make the following statement for the required complexity of a shallow network to approximate a function in a Sobolev space.

Let $1 \leq d \leq s$, $r \geq 1$, $n \geq 1$ be integers and $1 \leq p \leq \infty$. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be infinitely many times continuously differentiable in some open sphere in $\mathbb{R}^d$. It is further assumed that there exists a point b in this sphere such that

$$D^k \phi(b) \neq 0, \mathbf{k} \in \mathbb{Z}^d, \mathbf{k} \geq 0 \tag{1.1}$$

Note that the above condition, is equivalent to the statement that the function ϕ is not a polynomial. This is shown in [**?**].

Then, there exist $d \times s$ matrices $\{A_j\}_{j=1}^{n}$ with the following property: For any function f in the Sobolev space $W_{r,s}^p$, there exist coefficients $a_j(f)$ such that

$$\left\| f - \sum_{j=1}^{n} a_j(f) \phi(A_j(\cdot) + b_j) \right\|_p \leq c n^{-r/s} \|f\|_{W_{r,s}^p} \tag{1.2}$$

Here, the functionals a_j are continuous linear functionals on $W_{r,s}^p$.

In particular, we measure the *degree of approximation* of f by the expression

$$E_{\phi;n,p} = \inf\{\|f - g\|_p : g \in \Pi_{\phi;n,s}\}$$

The quantity $E_{\phi;n,p}$ denotes the theoretically minimal error that can be achieved in approximating the function f in the L^p norm by generalized translation networks with n neurons each evaluating the activation function ϕ.

$$E_{\phi;n,p,r,s} := \sup\{E_{\phi;n,p}(f) : \|f\|_{W_{r,s}^p} \leq 1\} \leq c n^{-r/s}$$

where $E_{\phi;n,r,s,p}$ denotes the maximal error that can be achieved in approximating functions in $W_{r,s}^p$ by generalized translation networks with n neurons each evaluating the activation function ϕ, with the assumption that the target function $f \in W_{r,s}^p$ is properly normalised. Here we take the fact that that any function in $W_{r,s}^p$ can be normalized so that $\|f\|_{W_{r,s}^p} \leq 1$.

1.2 Proof

1.2.1 Idea for Proof

For every integer $m \geq r, \exists$ polynomial $P_m(f)$ of co-ordinatewise degree not exceeding m such that $\forall f \in W^p_{r,s}$. By co-ordinatewise degree, we mean that the degree of the polynomial in each variable is at most m.

$$\|f - P_m(f)\|_{p,[-\pi,\pi]^s} \leq \frac{c}{m^r}\|f\|_{W^{p*}_{r,s}}$$

Aim to express each monomial in $P_m(f)$ with suitable derivative of ϕ, then take each derivative approximations and approximate it by an appropriate divided difference method involving $O(m^s)$ evaluations of ϕ

1.2.2 Useful Definitions and clarifications

- $|\mathbf{k}| = \sum_j k_j$
- $0 \leq \mathbf{k} \leq \mathbf{r} \iff 0 \leq k_j \leq r_j \forall j$
- $L^{p^*} := L^p([-\pi,\pi]^s)$, $W^{p^*}_{r,s} := W^p_{r,s}([-\pi,\pi]^s)$
- **Fourier Coefficients**: For a function $g \in L^{p^*} = L^p([-\pi,\pi]^s)$, its Fourier coefficients are given by the following, where $\mathbf{k}$ is a multi-index in $\mathbb{Z}^s$, and $\mathbf{t}$ is in the domain $[-\pi,\pi]^s$.

$$g(\mathbf{k}) := \frac{1}{(2\pi)^s}\int_{[-\pi,\pi]^s} g(\mathbf{t})e^{-i\mathbf{kt}}d\mathbf{t}, \quad \mathbf{k} \in \mathbb{Z}^s. \tag{1.3}$$

- **Partial Sums of the Fourier Series**: The partial sum $s_m(g,t)$ of the Fourier series of g is given as follows,

$$s_m(g(\mathbf{k}),t) := \sum_{-\mathbf{m}\leq\mathbf{k}<\mathbf{m}} g(\mathbf{k})e^{i\mathbf{kt}}, \quad \mathbf{m} \in \mathbb{Z}^s, \quad \mathbf{m} \geq 0, \quad \mathbf{t} \in [-\pi,\pi]^s, \tag{1.4}$$

- **de la Vallée Poussin Operator**: The de la Vallée Poussin operator $v_n(g,t)$ for a function g is defined as the average of the partial sums $s_m(g,t)$ where m ranges from n to $2n$. This operator is used to create a smoothed approximation of the function g.

$$v_n(g,\mathbf{t}) := \frac{1}{(n+1)^s}\sum_{n\leq\mathbf{m}\leq 2n} s_{\mathbf{m}}(g,t), \quad n \in \mathbb{Z}, \quad n \geq 0, \quad \mathbf{t} \in [-\pi,\pi]^s, \tag{1.5}$$

We occasionally omit the $\mathbf{k}$ argument of $g(\mathbf{k})$ for brevity.

Proposition 1.1 ([?]). *We have the following result:*

- *Given integers $r \geq 1$, s, $m \geq 1$, and $1 \leq p \leq \infty$.*
- *Let g be a function in the Sobolev space $W^{p^*}_{r,s}$.*
- *Then $v_m(g)$ is defined as a trigonometric polynomial of coordinate-wise order at most $2m$. This means that $v_m(g)$ is a polynomial composed of sine and cosine terms, where the highest frequency term has a frequency of $2m$.*

  Here $v_m(g)$ is defined as the de la Vallée Poussin operator applied to g, we omit the t dependence for brevity. $v_m(g) \equiv v_m(g(\mathbf{k}),\mathbf{t}) \equiv v_m(g,\mathbf{t})(\mathbf{k})$

- *And*

$$\|g - v_m(g)\|_{p,[-\pi,\pi]^s} \leq \frac{c}{m^r}\|g\|_{W^{p*}_{r,s}} \tag{1.6}$$

- *further,*

$$\sum_{0\leq\mathbf{k}<2m} |\hat{v_m(g)}(\mathbf{k})| \leq cm^\alpha \|g\|_{W^{p*}_{r,s}} \quad \textit{where} \quad \alpha = \frac{s}{\min(p,2)} \tag{1.7}$$

The idea follows to make a periodic function from a function on $[-1,1]^s$. The standard way to achieve this is via a cosine substitution; $x_j = \cos(t_j) 1 \leq j \leq s, \mathbf{x} \in [-1,1]^s, \mathbf{t} \in [-\pi,\pi]^s$. This substitution doesn't work as the L^p norms are no longer equal under this substitution.

Instead construct the following.
We first have from [**?**] that there exists a continuous linear operator $T : W^p_{r,s} \to W^p_{r,s}([-2,2]^s)$ such that the restriction of $T(f)$ on $[-1,1]^s$ is (almost everywhere) equal to f. The continuity of the operator T means that

$$\|T(f)\|_{W^p_{r,s}([-2,2]^s)} \leq c\|f\|_{W^p_{r,s}}$$

for every $f \in W^p_{r,s}$, This means that the operator T maps a function in the Sobolev space $W^p_{r,s}$ to a function in the same space but defined on a larger domain, $[-2,2]^s$, we construct this operator later.

If we find in practice we have an f already appropriately defined on $[-2,2]^s$, then we can simply work with f itself rather than $T(f)$, however our bounds would then depend on the respective Sobolev space; $W^p_{r,s}([-2,2]^s)$

We now define ψ be an infinitely differentiable function that takes the value 1 on $[-1,1]^s$ and 0 outside the interval $[-3/2,3/2]^s$. Then the function $T(f)\psi$ coincides with f on $[-1,1]^s$ and is identically zero outside $[-3/2,3/2]^s$ and

$$\|T(f)\psi\|_{W^p_{r,s}([-2,2]^s)} \leq c\|f\|_{W^p_{r,s}} \tag{1.8}$$

We relabel $T(f)\psi$ as f for brevity. We then define a 2π-periodic function f^* from the function f (extended as above) using the transformation $x_j = 2\cos(t_j) 1 \leq j \leq s, \mathbf{x} \in [-1,1]^s, \mathbf{t} \in [-\pi,\pi]^s$.

$$f^*(\mathbf{t}) = f(\mathbf{x}) = f(2\cos(t_1), ..., 2\cos(t_s)) \quad \mathbf{t} \in [-\pi,\pi]^s \tag{1.9}$$

We find that $f^* \in W^{p^*}_{r,s}$. Using induction and the fact that f is identically zero outside $[-3/2,3/2]^s$, using the previous bound (1.8) we find

$$c_1\|f\|_{W^p_{r,s}} \leq \|f^*\|_{W^{p^*}_{r,s}} \leq c_2\|f\|_{W^p_{r,s}} \tag{1.10}$$

We can now check that for any integer m, $v_m(f^*)$ is even and can be written as a sum of just cosines. We write for some set of coefficients $V_\mathbf{k}(f)$, the de la Vallée Poussin operator applied to f^* as

$$v_m(f^*, \mathbf{t}) = \sum_{0\leq\mathbf{k}\leq 2m} V_\mathbf{k}(f) \prod_{j=1}^{s} \cos(k_j t_j) \tag{1.11}$$

For integer $k \geq 0$, we let T_k be the Chebyshev polynomial adapted to the interval $[-2,2]$ as defined by

$$T_k(2\cos t) = \cos(kt) \quad \text{for} \quad t \in [-\pi,\pi] \tag{1.12}$$

And for a multi-integer $\mathbf{k} \geq 0$, let

$$T_{\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{s} T_{k_j}(x_j) \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^s \tag{1.13}$$

The polynomial $P_m(f)$ of co-ordinatewise degree at most $2m$ is then defined as

$$P_m(f, \mathbf{x}) = \sum_{0 \leq \mathbf{k} \leq 2m} V_{\mathbf{k}}(f) T_{\mathbf{k}}(\mathbf{x}) \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^s \tag{1.14}$$

We relate $P_m(f)$ to $v_m(f^*)$ by the following

$$P_m(f, (2\cos t_1, ..., 2\cos t_s)) = v_m(f^*, \mathbf{t}) \quad \text{for} \quad \mathbf{t} \in [-\pi, \pi]^s \tag{1.15}$$

Going back to our key result and equation (1.6). We have that

$$\|f - P_m(f)\|_{p,[-\pi,\pi]^s} \leq \frac{c}{m^r} \|f\|_{W^{p*}_{r,s}} \tag{1.16}$$

and that

$$\sum_{0 \leq \mathbf{k} \leq 2m} |V_{\mathbf{k}}(f)| \leq c m^{\alpha} \|f\|_{W^{p*}_{r,s}} \quad \text{where} \quad \alpha = \frac{s}{\min(p, 2)} \tag{1.17}$$

To finish our proof we now have to construct an approximation to every polynomial. This is achieved by the following lemma.

Lemma 1.2. *States that given a function ϕ satisfying conditions from Theorem 1, for any integer $m \geq 1$ and any multi-integer $\mathbf{k} \in \mathbb{Z}^s$ with each component $\max_{1 \leq j \leq s} |k_j| \leq m$, there exists a function $G_{\mathbf{k},m,\epsilon} \in \Pi_{\phi;(6m+1)^s,s}$ such that*

$$\|T_{\mathbf{k}} - G_{\mathbf{k},m,\epsilon}\|_{\infty} \leq \epsilon \tag{1.18}$$

The weights and thresholds of each $G_{\mathbf{k},m,\epsilon}$ may be chosen from a fixed set with cardinality not exceeding $(6m+1)^s$

Proof. First consider the case when $d = 1$. We take the point $\mathbf{b}$ from (1.1), for $d = 1$ this is a real number, $\mathbf{b} = b$. Let ϕ be infinitely many times continuously differentiable on $[b - \delta, b + \delta]$

For a multi-integer $\mathbf{p} = (p_1, \ldots, p_s)$ and $\mathbf{x} \in \mathbb{R}^s$ we write

$$\mathbf{x}^{\mathbf{p}} := \prod_{j=1}^{s} x_j^{p_j} \quad \text{where we take } 0^0 = 1 \tag{1.19}$$

We see from

$$\phi_p(\mathbf{w}; \mathbf{x}) := \frac{\partial^{|p|}}{\partial w_{p_1} \ldots \partial w_{p_s}} \phi(\mathbf{w} \cdot \mathbf{x} + b) = x^{\mathbf{p}} \phi^{(|\mathbf{p}|)}(\mathbf{w} \cdot \mathbf{x} + b), \tag{1.20}$$

we conclude that

$$\mathbf{x}^{\mathbf{p}} = \left(\phi^{(|\mathbf{p}|)} \mathbf{w}(b)\right)^{-1} \phi_{\mathbf{p}}(\mathbf{0}; x) \tag{1.21}$$

We now apply our appropriate divided difference to replace $\phi_{\mathbf{p}}(\mathbf{0}; x)$. For multi-integers $\mathbf{p}$ and $\mathbf{r}$, we write

$$\binom{p}{r} := \prod_{j=1}^{s} \binom{p_j}{r_j}. \tag{1.22}$$

For any $h > 0$, the network defined by the formula

$$\Phi_{p,h}(\mathbf{x}) := h^{-|p|} \sum_{0 \leq r \leq p} (-1)^{|r|} \binom{p}{r} \phi(h(2r - p) \cdot \mathbf{x} + b) \tag{1.23}$$

is in $\Pi_{\phi;(p_1+1),\ldots,(p_s+1)}$, and represents a divided difference for $\phi_p(0;\mathbf{x})$. Further, we have

$$\|\Phi_{p,h} - \phi_p(0;\cdot)\|_\infty \le M_{\phi,m,s} h^2, \quad \max_{1\le j\le s} |\, p_j \,| \le m, |\, h \,| \le \delta/(3ms) \tag{1.24}$$

where $M_{\phi,m,s}$ is a positive constant depending only on the indicated variables.
Now, we write $T_k(\mathbf{x}) := \sum_{0\le p\le k} \tau_{\mathbf{k},\mathbf{p}} \mathbf{x}^{\mathbf{P}}$, and choose

$$h := h_{\phi,m,s} := \min\left\{\delta, \min_{0\le s\le 2m}\left(\frac{\epsilon}{M_{\phi,m,s}\sum_{0\le p\le k}|\phi(b)|^{-1}|\tau_{k,p}|}\right)^{1/2}\right\}. \tag{1.25}$$

Then the above equation implies that the network $G_{k,m,\epsilon}$ defined by

$$G_{k,m,\epsilon}(\mathbf{x}) := \sum_{0\le p\le k} \tau_{k,p}(\phi(b))^{-1}\Phi_{p,h_{\phi,m,s}}(\mathbf{x}), \tag{1.26}$$

satisfies the bound given in the lemme. For each $\mathbf{k}$, the weights and thresholds in $G_{k,m,\epsilon}$ are chosen from the set

$$\{(h_{\phi,m,s}, b) : b \in \mathbb{Z}^s, |r_j| \le 3m, 1 \le j \le s\}. \tag{1.27}$$

The cardinality of this set is $(6m+1)^s$. Therefore, $G_{k,m,\epsilon} \in \Pi_{\phi;(6m+1)^s}$.
Next, if $d > 1$, and $\mathbf{b}$ is as in the original theorem, then we consider the univariate function

$$\sigma(x) := \phi(x, b_2, \ldots, b_d) \tag{1.28}$$

The function σ satisfies all the hypothesis of Theorem 1, with b_1 in place of $\mathbf{b}$. Taking into account the fact that $\sigma(\mathbf{w}\cdot\mathbf{x} + b_1) = \phi(A_{\mathbf{w}}\mathbf{x} + b)$ with

$$A_{\mathbf{w}} := \begin{pmatrix} \mathbf{w} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{1.29}$$

any network in $\Pi_{\sigma;n,s}$ is also a network in $\Pi_{\phi;n,s}$. Therefore, the case $d = 1$ implies the lemma also when $d > 1$. □

Theorem 1 Proof. Without loss of generality, we may assume that $n \ge 13^s$. Let $m \ge 1$ be the largest integer such that $(12m+1)^s \le n$. We define $P_m(f) = \sum_{0\le k\le 2m} V_k(f)T_k$ as before. We then define the network

$$N_n(f,x) := \sum_{0\le k\le 2m} V_k(f) G_{k,2m,m-r-\alpha}(x) \tag{1.30}$$

is in $\Pi_{\phi;n,s}$ and satisfies

$$\|P_m(f) - N_n(f)\|_\infty \le cm^{-r}\|f\|_{W^p_r,s}. \tag{1.31}$$

Since $\|g\|_p \le 2^{s/p}\|g\|_\infty$ for all Lebesgue measurable functions g on $[-1,1]$ we get from our key result (Proposition 1.1, equation 1.6) that

$$\|f - N_n(f)\|_p \le cn^{-r/s}\|f\|_{W^p_r,s} \tag{1.32}$$

as required. Further, it is quite clear that the coefficients V_k are continuous linear functionals on L^p. Hence, the continuity assertion follows. □