

Approximation Theory and Neural Networks

H. N. Mhaskar*

Department of Mathematics, California State University
Los Angeles, California, 90032, U.S.A.

1 Approximation by trigonometric polynomials

1.1 Introduction

In many practical situations, one needs to construct a model for an input/output process. For example, one is interested in the price of a stock five years from now. The rating industry description for the stock typically lists such indicators as the increase in the price over the last year, the last 5 years, 10 years, life of the stock, P/E ratio, and alpha and beta risk factors. The buyer is expected to (but instructed not to!) believe that the price of the stock depends upon these parameters. Of course, no one knows a precise formula to compute this price as a closed form function of the parameters, but only has available data on the many stocks traded on the market.

The general situation is as in Figure 1. In general, the model P_f has to be constructed

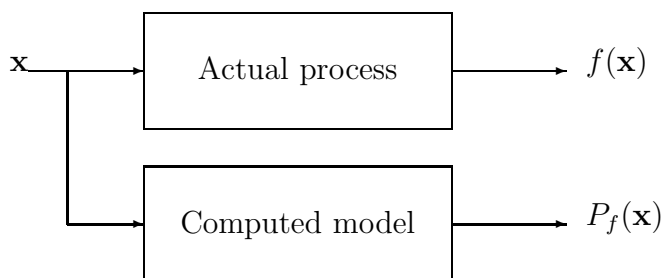


Figure 1: The function P_f represents a model for the actual function f of the input \mathbf{x} in some Euclidean space.

*The research of this author was supported, in part, by grant DMS-9971846 from the National Science Foundation, and by grant F-49620-97-1-0211 from the U. S. Air Force Office of Scientific Research.

based on finitely many observations on the *target function* f . Quite often, these are just the values of f at certain points, but in different applications, they may be the values of the derivatives, or Fourier coefficients of f , or the coefficients in some other series associated with f , etc.

There are two kinds of “errors” involved in using $P_f(\mathbf{x})$ as a predictor for $f(\mathbf{x})$. The *intrinsic error* arises from the fact that we are computing a model rather than the actual function. The second, often called *noise*, comes from the fact that the observations on which the model is based contain errors. The noise comes, for example, from human errors, instrumentation errors, interference from “nature”, and also, as in the case of the stock price example above, from our assumptions about what we are modelling. The subject of statistics deals with the problem of making a model reliable by eliminating or controlling the noise. The subject of approximation theory deals with the intrinsic error.

The problems of approximation theory can be grouped loosely in four categories. For explaining these, we take the example where the target function $f : \mathbb{R} \rightarrow \mathbb{R}$ is a continuous 2π -periodic function. The *norm* of f is defined by

$$\|f\|^* := \max_{x \in [-\pi, \pi]} |f(x)|.$$

The model for such a function is typically a trigonometric polynomial; i.e., an expression of the form $\sum_{|k| \leq n} c_k e^{ikx}$, $c_k \in \mathbb{C}$, $k = 0, \pm 1, \pm 2, \dots$. If $|c_n| + |c_{-n}| \neq 0$, the integer n is called the order of the polynomial. The class of all trigonometric polynomials of order at most n is denoted by \mathbb{H}_n .

The *density problem* consists of deciding whether it is possible to approximate the target function arbitrarily well by choosing more and more complex models. In our example, the *degree of approximation* of f from \mathbb{H}_n is defined by

$$E_n^*(f) := \inf_{T \in \mathbb{H}_n} \|f - T\|^*.$$

The density problem is to decide if $E_n^*(f) \rightarrow 0$ as $n \rightarrow \infty$. (It does; because of the Fejér-Weierstrass theorem. See remarks following (1.2.9).)

The *complexity problem* deals with estimating the rate at which $E_n^*(f) \rightarrow 0$. We observe that the target function is typically unknown. Therefore, in theoretical considerations, one makes some assumptions about the function, for example, that it has a continuous derivative, whose norm is bounded by 1. These assumptions are encoded by the statement that $f \in W$ for some function class W . Since the target function itself is unknown, one is interested in estimating

$$\sup_{f \in W} E_n^*(f)$$

as a function of n .

The *theory of best approximation* deals with the existence and properties of $T^* \in \mathbb{H}_n$ such that

$$\|f - T^*\|^* = E_n^*(f).$$

The *theory of good approximation*, on the other hand, deals with the approximation capabilities of different procedures to compute approximating trigonometric polynomials

based on the values of the function and its derivatives, its Fourier coefficients, etc. These approximants are sometimes more interesting than the best approximant, T^* , because they are easier to compute than the best approximant, or because they possess certain desirable properties, for example, shape preservation, that are not shared by T^* .

Approximation theory has widely influenced such other areas of mathematics as orthogonal polynomials, partial differential equations, harmonic analysis, wavelet analysis. Some modern applications include computer graphics, signal processing, economic forecasting, and pattern recognition.

In the next five lectures, I will focus on certain aspects of the complexity problem and the theory of good approximation, particularly as they apply to neural networks. In the first three lectures, I will discuss these issues as they relate to approximation of periodic functions by trigonometric polynomials. In the fourth lecture, I will discuss the connection between neural networks and trigonometric approximation, and in the fifth, the connection between neural networks and algebraic polynomial approximation.

1.2 Favard inequality

In this section, we are interested in the approximation of continuously differentiable 2π -periodic functions. Let C^* denote the class of all 2π -periodic continuous functions on \mathbb{R} , and W_r^* denote the class of all r times continuously differentiable 2π -periodic functions on \mathbb{R} . If $\rho > 0$, $\rho = r + \alpha$ for some integer $r \geq 0$ and $\alpha \in (0, 1)$, we say that $f \in \Lambda_\rho^*$ if $f \in W_r^*$ and

$$\|f\|_\rho^* := \sup_{\substack{x, y \in [-\pi, \pi] \\ x \neq y}} \frac{|f^{(r)}(x) - f^{(r)}(y)|}{|x - y|^\alpha} < \infty. \quad (1.2.1)$$

Although it is not a standard practice, for the simplicity of exposition, we will make the convention that for integer $\rho \geq 1$, $\Lambda_\rho^* = W_\rho^*$ and

$$\|f\|_\rho^* := \|f^{(\rho)}\|^*. \quad (1.2.2)$$

Our main objective is to prove Theorem 1.2.1 given below. We find it convenient to adopt the following convention regarding constants. The symbols c, c_1, \dots denote positive constants depending on the fixed parameters of the problem, such as r , but independent of f and n . Their value may be different at different occurrences, even within the same formula.

Theorem 1.2.1 *Let $\rho > 0$ and $f \in \Lambda_\rho^*$. For integers $n = 1, 2, \dots$, we have*

$$E_n^*(f) \leq cn^{-\rho} \|f\|_\rho^*. \quad (1.2.3)$$

In particular, if $r \geq 1$ is an integer,

$$E_n^*(f) \leq cn^{-r} \|f^{(r)}\|^*. \quad (1.2.4)$$

In order to prove this theorem, we introduce some notations. For a 2π -periodic integrable function f , and integers $k \in \mathbb{Z}$, $m, n \geq 1$, we write

$$\begin{aligned} c_k^*(f) &:= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-ikt} dt, \\ s_m^*(f, x) &:= \sum_{|k| \leq m-1} c_k^*(f) e^{ikx}, \\ \sigma_n^*(f, x) &:= \frac{1}{n} \sum_{m=1}^n s_m^*(f, x). \end{aligned} \tag{1.2.5}$$

We will obtain an integral expression for s_m^* and σ_n^* . First, we observe that

$$\begin{aligned} &\sin(u/2) \sum_{|k| \leq m-1} e^{iku} \\ &= 2 \sin(u/2) \left(\frac{1}{2} + \sum_{k=1}^{m-1} \cos ku \right) \\ &= \sin(u/2) + \sum_{k=1}^{m-1} (\sin(k+1/2)u - \sin(k-1/2)u) \\ &= \sin(m-1/2)u, \end{aligned}$$

and hence,

$$D_m^*(u) := \sum_{|k| \leq m-1} e^{iku} = \frac{\sin(m-1/2)u}{\sin(u/2)}. \tag{1.2.6}$$

Using a similar argument, it is easy to verify that

$$F_n^*(u) := \frac{1}{n} \sum_{m=1}^n D_m^*(u) = \frac{1}{n} \left(\frac{\sin(nu/2)}{\sin(u/2)} \right)^2. \tag{1.2.7}$$

Therefore, using the first two equations in (1.2.5), we obtain

$$\begin{aligned} s_m^*(f, x) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \left(\sum_{|k| \leq m-1} e^{ik(x-t)} \right) dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) D_m^*(x-t) dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-t) D_m^*(t) dt. \end{aligned} \tag{1.2.8}$$

Similarly,

$$\sigma_n^*(f, x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) F_n^*(x-t) dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-t) F_n^*(t) dt. \tag{1.2.9}$$

It is well known that there exists $f \in C^*$ for which $s_m^*(f)$ does not converge uniformly to f , while the Fejér-Weierstrass theorem asserts that $\sigma_n^*(f)$ converges uniformly to f for every $f \in C^*$. However, the rate of convergence is not what is required by Theorem 1.2.1.

To accomplish this rate, we introduce a modification of F_n^* , called the *Jackson kernel*. For integer $n \geq 1$, we define

$$K_n^*(u) := \left(\frac{\sin(nu/2)}{\sin(u/2)} \right)^4. \quad (1.2.10)$$

Since $F_n^* \in \mathbb{H}_n$, it is clear that $K_n^* \in \mathbb{H}_{2n}$. Denoting, in the remainder of this section only, the integer part of $n/2$ by n' , and

$$\lambda_n^{-1} := \frac{1}{2\pi} \int_{-\pi}^{\pi} K_{n'}^*(u) du,$$

we define

$$J_n(f; x) := \frac{\lambda_n}{2\pi} \int_{-\pi}^{\pi} f(t) K_{n'}^*(x-t) dt = \frac{\lambda_n}{2\pi} \int_{-\pi}^{\pi} f(x-t) K_{n'}^*(t) dt. \quad (1.2.11)$$

It is easy to verify that $J_n(f) \in \mathbb{H}_n$. We summarize certain properties of the operator J_n^* . In the sequel, we write $A \sim B$ to denote the fact that $cA \leq B \leq c_1A$.

Lemma 1.2.1 *For $0 \leq \alpha \leq 2$ and $n = 1, 2, 3, \dots$, we have*

$$\int_{-\pi}^{\pi} |u|^\alpha K_n^*(u) du \sim n^{3-\alpha}. \quad (1.2.12)$$

In particular,

$$\lambda_n \leq cn^{-3}, \quad \int_{-\pi}^{\pi} |u| K_{n'}^*(u) du \leq cn^2. \quad (1.2.13)$$

PROOF. Using the mean value theorem and the fact that the function $x \mapsto \sin(x/2)$ is concave on $[0, \pi]$, we obtain that

$$\frac{t}{\pi} \leq \sin(t/2) \leq \frac{t}{2}, \quad t \in [0, \pi]. \quad (1.2.14)$$

So, with $t = nu$,

$$\begin{aligned} \int_{-\pi}^{\pi} |u|^\alpha K_n^*(u) du &= 2 \int_0^{\pi} u^\alpha K_n^*(u) du \\ &\sim \int_0^{\pi} \sin^4(nu/2) u^{\alpha-4} du = n^{3-\alpha} \int_0^{n\pi} \sin^4(t/2) t^{\alpha-4} dt. \end{aligned} \quad (1.2.15)$$

Now, since $\alpha - 4 \leq 0$,

$$\begin{aligned} 1 &= \frac{8}{3\pi} \int_0^{\pi} \sin^4(t/2) dt \\ &\leq \frac{8\pi^{3-\alpha}}{3} \int_0^{\pi} \sin^4(t/2) t^{\alpha-4} dt \\ &\leq \frac{\pi^3}{6} \int_0^{\pi} \left(\frac{\sin(t/2)}{t/2} \right)^4 dt \leq \frac{\pi^4}{6}. \end{aligned}$$

Also, since $\alpha - 4 \leq -2$,

$$\begin{aligned} 0 &\leq \int_{\pi}^{n\pi} \sin^4(t/2)t^{\alpha-4}dt \\ &\leq \int_{\pi}^{\infty} t^{\alpha-4}dt = \frac{\pi^{\alpha-3}}{3-\alpha}. \end{aligned}$$

Thus,

$$\frac{3}{8\pi^{3-\alpha}} \leq \int_0^{n\pi} \sin^4(t/2)t^{\alpha-4}dt \leq \frac{\pi^{1+\alpha}}{16} + \frac{\pi^{\alpha-3}}{3-\alpha}.$$

Along with (1.2.15), this leads to (1.2.12). \square

Proposition 1.2.1 For $0 < \rho \leq 1$, $f \in \Lambda_{\rho}^*$ and $n = 2, 3, \dots$, we have

$$\|f - J_n^*(f)\| \leq cn^{-\rho}\|f\|_{\rho}^*. \quad (1.2.16)$$

PROOF. Let $x \in [-\pi, \pi]$. Using (1.2.11) and the definition of λ_n , we obtain

$$\begin{aligned} f(x) - J_n^*(f, x) &= \frac{1}{2\pi}\lambda_n \int_{-\pi}^{\pi} f(x)K_{n'}^*(t)dt - \frac{1}{2\pi}\lambda_n \int_{-\pi}^{\pi} f(x-t)K_{n'}^*(t)dt \\ &= \frac{1}{2\pi}\lambda_n \int_{-\pi}^{\pi} (f(x) - f(x-t))K_{n'}^*(t)dt. \end{aligned}$$

Since $f \in \Lambda_{\rho}^*$, $|f(x) - f(x-t)| \leq |t|^{\rho}\|f\|_{\rho}^*$. (If $\rho < 1$, then this is clear from the definition. If $\rho = 1$, we use the mean value theorem.) Therefore, using (1.2.13), we obtain

$$|f(x) - J_n^*(f, x)| \leq cn^{-3}\|f\|_{\rho}^* \int_{-\pi}^{\pi} |t|^{\rho}K_{n'}^*(t)dt \leq cn^{-rho}\|f\|_{\rho}^*.$$

\square

Next, we establish a connection between $E_n^*(f)$ and $E_n^*(f^{(r)})$.

Proposition 1.2.2 Let $f \in W_r^*$. Then for integer $n \geq 1$,

$$E_n^*(f) \leq \frac{c}{n^r}E_n^*(f^{(r)}). \quad (1.2.17)$$

PROOF. We prove this proposition by induction on r . First, let $r = 1$, and $f \in W_1^*$. There exists $T_1 \in \mathbb{H}_n$ such that

$$\|f' - T_1\|^* \leq 2E_n^*(f').$$

Since $\int_{-\pi}^{\pi} f'(t)dt = 0$, we see that $|(1/(2\pi)) \int_{-\pi}^{\pi} T_1(t)dt| \leq 2E_n^*(f')$. Therefore,

$$T_2(x) := T_1(x) - \frac{1}{2\pi} \int_{-\pi}^{\pi} T_1(t)dt$$

satisfies

$$\|f' - T_2\|^* \leq 4E_n^*(f'). \quad (1.2.18)$$

Now,

$$R(x) := \int_{-\pi}^x T_2(t)dt \in \mathbb{H}_n,$$

and $(f - R)' = f' - T_2$. Therefore, by Proposition 1.2.1 and (1.2.18), the polynomial $T := R + J_n^*(f - R) \in \mathbb{H}_n$ satisfies

$$\|f - T\|^* = \|f - R - J_n^*(f - R)\|^* \leq \frac{c}{n} \|(f - R)'\|^* \leq \frac{c}{n} E_n^*(f').$$

This proves (1.2.17) in the case $r = 1$. We obtain the general case easily by a repeated application of the case $r = 1$. \square

PROOF OF THEOREM 1.2.1. The theorem follows immediately from Propositions 1.2.2 and 1.2.1. \square

1.3 Shifted average operator

During the proof of Theorem 1.2.1, we gave a recursive construction for the trigonometric polynomial that approximates the target function with the correct rate. However, since we don't know the target function, it is desirable to get some construction which works without knowing the number of derivatives of the target function in advance. The best approximation will work, but there is a substantially easier construction which gives a good approximation.

Theorem 1.3.1 *For an integrable, 2π -periodic function f , and integer $n \geq 1$, let*

$$v_n^*(f) := \frac{1}{n} \sum_{m=n+1}^{2n} s_m(f). \quad (1.3.1)$$

(a) *We have*

$$v_n^*(f, x) = s_{n+1}^*(f, x) + \sum_{k=n+1}^{2n-1} \left(2 - \frac{k}{n}\right) (c_k^*(f)e^{ikx} + c_{-k}^*(f)e^{-ikx}). \quad (1.3.2)$$

In particular, $v_n^(T) = T$ for all $T \in \mathbb{H}_n$.*

(b) *With*

$$V_n^*(u) := \frac{\sin(nu/2) \sin(3nu/2)}{n \sin^2(u/2)}, \quad (1.3.3)$$

$$v_n^*(f, x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) V_n^*(x - t) dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x - t) V_n^*(t) dt. \quad (1.3.4)$$

(c) *For $f \in C^*$, we have $v_n^*(f) \in \mathbb{H}_{2n-1}$, and*

$$\|v_n^*(f)\|^* \leq 3\|f\|^*, \quad E_{2n-1}^*(f) \leq \|f - v_n^*(f)\|^* \leq 4E_n^*(f). \quad (1.3.5)$$

(d) *We have*

$$\|V_n^*\|^* \leq 3n, \quad \int_{-\pi}^{\pi} |V_n^*(u)| du \leq 6\pi. \quad (1.3.6)$$

PROOF. The proof of parts (a) and (b) are simple computations, using the definition of s_m^* and interchange of summation for part (a), and the integral representation (1.2.8) for part (b). From the definitions, we see that $v_n^*(f) = 2\sigma_{2n}^*(f) - \sigma_n^*(f)$. Now, (1.2.8) implies that $\|\sigma_n^*(f)\|^* \leq \|f\|^*$ for all n ; which implies the first estimate in (1.3.5). Since $v_n^*(f) \in \mathbb{H}_{2n-1}$, it is clear that $E_{2n-1}^*(f) \leq \|f - v_n^*(f)\|^*$. If $T \in \mathbb{H}_n$, then $v_n^*(T) = T$, and

$$\|f - v_n^*(f)\|^* = \|(f - T) - v_n^*(f - T)\|^* \leq \|f - T\|^* + \|v_n^*(f - T)\|^* \leq 4\|f - T\|^*.$$

Since the left-most term of this estimate is independent of T , we may take the infimum over all $T \in \mathbb{H}_n$ to arrive at the last estimate in (1.3.5). The first estimate in (1.3.6) is clear from (1.3.3). We get the second estimate in (1.3.6) from the observations that $V_n^* = 2F_{2n}^* - F_n^*$, $F_m^*(u) \geq 0$ ($u \in [-\pi, \pi]$), and $\int_{-\pi}^{\pi} F_m^*(u) du = 2\pi$ for $m = 1, 2, \dots$. \square

1.4 Multivariate extension

Let $s \geq 1$ be an integer. The class of all continuous functions $f : \mathbb{R}^s \rightarrow \mathbb{R}$ which are 2π -periodic in each of the variables is denoted by C_s^* . For $f \in C_s^*$, we write

$$\|f\|_s^* := \max_{\mathbf{x} \in [-\pi, \pi]^s} |f(\mathbf{x})|.$$

The class of all functions in C_s^* having continuous partial derivatives up to order r in each variable is denoted by $W_{r,s}^*$. As usual, the notation $D_j f$ denotes the partial derivative of f with respect to its j -th variable, and for $\mathbf{k} = (k_1, \dots, k_s) \in \mathbb{Z}^s$, $\mathbf{k} \geq 0$ (i.e., $k_j \geq 0$ for $1 \leq j \leq s$), $D^{\mathbf{k}} f$ denotes $D_s^{k_s} \dots D_1^{k_1} f$. The class of all trigonometric polynomials in s variables will be denoted by $\mathbb{H}_{n,s}$. Finally, for $f \in C_s^*$, we write

$$E_{n,s}^*(f) := \inf_{T \in \mathbb{H}_{n,s}} \|f - T\|_s^*.$$

All constants in this section will depend upon r and s .

Theorem 1.4.1 *Let $r \geq 1$ be an integer. For integer $n \geq 1$ and $f \in W_{r,s}^*$, we have*

$$E_{n,s}^*(f) \leq \frac{c}{n^r} \sum_{j=1}^s \|D_j^r f\|_s^*. \quad (1.4.1)$$

PROOF. In this proof, we write

$$\begin{aligned} v_{n,j}^*(f, \mathbf{x}) &:= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x_1, \dots, x_{j-1}, x_j - t, x_{j+1}, \dots, x_s) V_n^*(t) dt, \\ v_n^{[j]}(f) &:= v_{n,j}^*(\dots(v_{n,1}^*(f))\dots), \quad v_n^{[0]}(f) := f. \end{aligned}$$

In view of (1.3.5), we get for $j = 1, \dots, s$, $\|v_n^{[j]}(D_j^r f)\|_s^* \leq 3^j \|D_j^r f\|_s^*$, and hence, using (1.2.4),

$$\begin{aligned} \|v_n^{[j]}(f) - v_n^{[j-1]}(f)\|_s^* &= \|v_{n,j}^*(v_n^{[j-1]}(f)) - v_n^{[j-1]}(f)\|_s^* \\ &\leq \frac{c}{n^r} \|D_j^r v_n^{[j-1]}(f)\|_s^* = \frac{c}{n^r} \|v_n^{[j-1]}(D_j^r f)\|_s^* \\ &\leq \frac{c}{n^r} \|D_j^r f\|_s^*. \end{aligned}$$

So, for the polynomial $v_n^{[s]}(f) \in \mathbb{H}_{2n-1,s}$, we have

$$\|v_n^{[s]}(f) - f\|_s^* \leq \sum_{j=1}^s \|v_n^{[j]}(f) - v_n^{[j-1]}(f)\|_s^* \leq \frac{c}{n^r} \sum_{j=1}^s \|D_j^r f\|_s^*.$$

□

We remark that the dimension of $\mathbb{H}_{n,s}$ is $\mathcal{O}(n^s)$. Hence, the number of parameters involved in approximating an arbitrary function in $W_{r,s}^*$ within an error of $\epsilon > 0$ is $\mathcal{O}(\epsilon^{-r/s})$.

2 Widths and converse theorems

2.1 Introduction

In this chapter, we show that the results of the previous chapters are sharp. We do this in two ways. We will show that no “reasonable approximation method” based on the same number of parameters can do asymptotically better than \mathbb{H}_n for the whole class W_r^* . We will also show that if the degree of approximation is as in (1.2.3) for some noninteger ρ , then the target function belongs to W_ρ^* . For simplicity, we will limit our discussion to the univariate case. The multivariate case does not offer any new features in this context.

Our proof of both of these results require an interesting fact about trigonometric polynomials, known as the Bernstein inequality. In the next section, we prove this inequality. In Section 2.3, we explain our first remark about “reasonable approximation methods” by introducing the notion of nonlinear widths, and obtaining a lower estimate for the widths for W_r^* . In Section 2.4, we obtain the converse of the estimate (1.2.3).

2.2 Bernstein inequality

In this section, our main objective is to prove the following theorem. For future reference, we prove it in greater generality than that required in this chapter.

Theorem 2.2.1 *Let $r, n \geq 1$ be integers, $T \in \mathbb{H}_n$, $p \geq 1$. Then*

$$\|T^{(r)}\|_*^* \leq n^r \|T\|_*^*, \tag{2.2.1}$$

and

$$\int_{-\pi}^{\pi} |T^{(r)}(t)|^p dt \leq n^{rp} \int_{-\pi}^{\pi} |T(t)|^p dt. \tag{2.2.2}$$

In this section only, we write $u_{j,n} := (2j - 1)\pi/(2n)$, $j = 1, \dots, 2n$, $n = 1, 2, \dots$, and for an integrable 2π -periodic function f , $a_k^*(f) := c_k^*(f) + c_{-k}^*(f)$, $k = 1, 2, \dots$.

Lemma 2.2.1 *Let $n \geq 1$ be an integer,*

$$D_n^{**}(u) := \frac{1}{2} + \sum_{k=1}^{n-1} \cos ku + \frac{1}{2} \cos nu, \quad u \in [-\pi, \pi]. \quad (2.2.3)$$

Then

$$D_n^{**}(u) = \frac{\sin nu}{2 \tan(u/2)}. \quad (2.2.4)$$

Every $S \in \mathbb{H}_n$ can be expressed in the form

$$S(x) = a_n^*(S) \cos nx + \frac{1}{n} \sum_{j=1}^{2n} S(u_{j,n}) D_n^{**}(x - u_{j,n}), \quad x \in [-\pi, \pi]. \quad (2.2.5)$$

PROOF. Using (1.2.6), we see that

$$1 + 2 \sum_{k=1}^{n-1} \cos ku + \cos nu = \frac{\sin(n-1/2)u}{\sin(u/2)} + \cos nu.$$

This leads to formula (2.2.4).

Next, we observe that for $k = 0, \pm 1, \pm 2, \dots, \pm 2n$,

$$\frac{1}{2n} \sum_{j=1}^{2n} \exp(iu_{j,n}k) = \begin{cases} 1, & \text{if } k = 0, \\ -1, & \text{if } k = \pm 2n, \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\frac{1}{2n} \sum_{j=1}^{2n} S(u_{j,n}) \exp(-iu_{j,n}k) = \begin{cases} c_k^*(S), & \text{if } |k| \leq n-1, \\ c_n^*(S) - c_{-n}^*(S), & \text{if } k = n, \\ c_{-n}^*(S) - c_n^*(S), & \text{if } k = -n, \end{cases}$$

A little computation now leads to (2.2.5). \square

PROOF OF THEOREM 2.2.1. We obtain a formula similar to (2.2.5) for T' . Let $S \in \mathbb{H}_n$. From (2.2.5), we obtain

$$S'(x) = -na_n^*(S) \sin nx + \frac{1}{2n} \sum_{j=1}^{2n} S(u_{j,n}) \times \left\{ n \cos n(x - u_{j,n}) \cot((x - u_{j,n})/2) - \frac{1}{2} \csc^2((x - u_{j,n})/2) \sin n(x - u_{j,n}) \right\}.$$

Therefore,

$$S'(0) = \frac{1}{n} \sum_{j=1}^{2n} S(u_{j,n}) \frac{(-1)^{j+1}}{(2 \sin(u_{j,n}/2))^2}. \quad (2.2.6)$$

Using this formula with $S(x) = T(x + t)$, we obtain the *Riesz representation formula*:

$$T'(t) = \frac{1}{n} \sum_{j=1}^{2n} T(t + u_{j,n}) \frac{(-1)^{j+1}}{(2 \sin(u_{j,n}/2))^2}, \quad t \in [-\pi, \pi]. \quad (2.2.7)$$

The formula (2.2.6) applied with $S(x) = \sin nx$ yields

$$\frac{1}{n} \sum_{j=1}^{2n} \left| \frac{(-1)^{j+1}}{(2 \sin(u_{j,n}/2))^2} \right| = \frac{1}{n} \sum_{j=1}^{2n} \frac{1}{(2 \sin(u_{j,n}/2))^2} = n. \quad (2.2.8)$$

For $r = 1$, (2.2.1) follows immediately from (2.2.7), and (2.2.8). The formula (2.2.2) for $r = 1$ also follows from (2.2.7), (2.2.8) if we use the Minkowski inequality. The formulas in the general case are proved by a repeated application of the inequalities in the case $r = 1$. \square

2.3 Widths

It is convenient to introduce the notion of nonlinear widths in the more general context of normed linear spaces. Let X be a normed linear space (e.g., C^*), and $K \subset X$ (e.g., W_r^*). Any method of approximation of elements of K using n parameters can be described as a composition of two mappings : the feature selection map $\phi : K \rightarrow \mathbb{R}^n$ (e.g., if n is odd, the mapping of a function f in C^* to $(c_{-m}^*(f), \dots, c_m^*(f))$, with $m = (n - 1)/2$) and a reconstruction map $\mathcal{M} : \mathbb{R}^n \rightarrow X$ (e.g., the shifted average operator defined using only the coefficients, without reference to the underlying function). The approximation to any $f \in K$ is given by $\mathcal{M}(\phi(f))$. The *worst case error* in this approximation is $\sup_{f \in K} \|f - \mathcal{M}(\phi(f))\|_X$, where $\|\cdot\|_X$ is the norm on X . The *nonlinear n -width* of K is defined by

$$\delta_n(K) := \delta_n(K, X) := \inf \sup_{f \in K} \|f - \mathcal{M}(\phi(f))\|_X, \quad (2.3.1)$$

where the infimum is taken over all $\mathcal{M} : \mathbb{R}^n \rightarrow X$ and *continuous* functions $\phi : K \rightarrow \mathbb{R}^n$. If K is not compact, then the definition depends upon the topology on K . The continuity of ϕ is required to avoid certain trivialities. In practice, it is a natural requirement in order to make sure that our parameter selection is stable under noise. The mapping of a function to all of its Fourier coefficients up to a certain order is continuous, the mapping to the greatest n Fourier coefficients is not.

The following theorem shows that the results of the previous chapters are optimal. In this section, B_r^* denotes the class of all $f \in W_r^*$ such that $\|f^{(r)}\|^* \leq 1$.

Theorem 2.3.1 *Let $r, n \geq 1$ be integers. We have*

$$\delta_n(B_r^*) \geq 2^r (n + r + 2)^{-r}. \quad (2.3.2)$$

The proof of this theorem relies upon the following theorem, known as the *Borsuk antipodality theorem*.

Theorem 2.3.2 *Let $(Y, \|\cdot\|_Y)$ be a finite dimensional normed linear space, $\rho > 0$, $S(Y, \rho) := \{y \in Y : \|y\|_Y = \rho\}$, $m \geq 1$ be an integer less than the dimension of Y , and $g : S(Y, \rho) \rightarrow \mathbb{R}^m$ be any continuous function. Then there exists $y_0 \in S(Y, \rho)$ such that $g(y_0) = g(-y_0)$.*

PROOF OF THEOREM 2.3.1. In this proof only, let

$$W := \{f \in B_r^* : f^{(j)}(0) = 0, j = 0, \dots, r-1\}.$$

Because of Arzela's theorem, W is a compact subset of C^* . Let $\phi : W \rightarrow \mathbb{R}^n$ be any continuous mapping, and $M : \mathbb{R}^n \rightarrow C^*$ be any function. Let, in this proof only, n' be the smallest integer with $2n' \geq n - r - 2$, and

$$Y := \{T \in \mathbb{H}_{n'+r+1} : T^{(j)}(0) = 0, j = 0, \dots, r-1\},$$

and $\|T\|_Y := \|T\|^*$. Then Y is a normed linear space of dimension $2n' + r + 3 > n$. In view of the Bernstein inequality, $S(Y, (n' + r + 1)^{-r}) \subset W$. The Borsuk antipodality theorem implies that there exists $T \in S(Y, (n' + r + 1)^{-r})$ such that $\phi(T) = \phi(-T)$. Then

$$2\|T\|^* = \|T - M(\phi(T)) - (-T - M(\phi(-T)))\|^* \leq \|T - M(\phi(T))\|^* + \|-T - M(\phi(-T))\|^*.$$

Hence, at least one of the inequalities $\|T - M(\phi(T))\|^* \geq \|T\|^*$ or $\|-T - M(\phi(-T))\|^* \geq \|-T\|^*$ must be true. In either case,

$$\sup_{f \in B_r^*} \|f - M(\phi(f))\|^* \geq \sup_{f \in S(Y, (n'+r+1)^{-r})} \|f - M(\phi(f))\|^* \geq \|T\|^* = (n' + r + 1)^{-r}.$$

Since $n' + r + 1 \leq (n - r)/2 + r + 1$, this completes the proof. \square

We remark that the ‘‘bad function’’ for which the lower estimate is true is actually a trigonometric polynomial $T \in \mathbb{H}_{(n+r+2)/2}$ with $\|T^{(r)}\|^* \leq 1$. The idea of the proof is applicable in general. One defines the *Bernstein n -width* of a subset K of a normed linear space by

$$\beta_n(K) := \sup_{Y \in \mathcal{X}_{n+1}} \sup\{\rho > 0 : S(Y, \rho) \subset K\}, \quad (2.3.3)$$

where we write \mathcal{X}_{n+1} to denote the class of all $n + 1$ dimensional subspaces of X , endowed with the norm of X . The same argument as in the proof of Theorem 2.3.1 then implies that

$$\delta_n(K) \geq \beta_n(K). \quad (2.3.4)$$

2.4 Converse theorems

The purpose of this section is to prove the following Theorem 2.4.1.

Theorem 2.4.1 *Let $\rho > 0$, $\rho \notin \mathbb{Z}$, and $f \in C^*$. Then the following are equivalent.*

- (a) *The function $f \in \Lambda_\rho^*$.*
- (b) *There exists a constant $B(f)$ such that*

$$E_n^*(f) \leq B(f)n^{-\rho}, \quad n = 1, 2, \dots. \quad (2.4.1)$$

PROOF. The implication (a) \Rightarrow (b) follows from Theorem 1.2.1. To prove the converse, suppose that (2.4.1) holds, and $\rho = r + \alpha$, $r \in \mathbb{Z}$, $\alpha \in (0, 1)$. First, we show that f has r continuous derivatives, and

$$E_n^*(f^{(r)}) \leq cB(f)n^{-\alpha}, \quad n = 1, 2, \dots \quad (2.4.2)$$

Let $n \geq 1$ be fixed. We may find a sequence of polynomials $R_k \in \mathbb{H}_{2^k n}$ such that $\|f - R_k\|^* \leq 2E_{2^k n}^*(f)$. Then the polynomials $T_k := R_{k+1} - R_k \in \mathbb{H}_{2^{k+1}n}$ and satisfy

$$\|T_k\|^* \leq 4E_{2^k n}^*(f) \leq 4B(f)n^{-\rho}2^{-k\rho}, \quad k = 0, 1, 2, \dots, \quad (2.4.3)$$

and (in the sense of uniform convergence)

$$f = R_0 + \sum_{k=0}^{\infty} T_k. \quad (2.4.4)$$

The Bernstein inequality implies that

$$\|T_k^{(r)}\|^* \leq 2^{(k+1)r} n^r \|T_k\|^* \leq cn^{-\alpha} 2^{-k\alpha} B(f), \quad k = 0, 1, 2, \dots \quad (2.4.5)$$

Hence, the series $R_0^{(r)} + \sum_{k=0}^{\infty} T_k^{(r)}$ converges uniformly. In view of (2.4.4), this implies that f has r continuous derivatives, and

$$f^{(r)} = R_0^{(r)} + \sum_{k=0}^{\infty} T_k^{(r)}$$

in the sense of uniform convergence. Since $R_0^{(r)} \in \mathbb{H}_n$, we conclude using (2.4.5) that

$$E_n^*(f^{(r)}) \leq \|f - R_0\|^* \leq \sum_{k=0}^{\infty} \|T_k^{(r)}\|^* \leq cn^{-\alpha} B(f). \quad (2.4.6)$$

This proves (2.4.2).

Thus, in order to complete the proof of the theorem, we need to show that if $g \in C^*$ and for some $\alpha \in (0, 1)$,

$$E_n^*(g) \leq B(g)n^{-\alpha}, \quad n = 1, 2, \dots, \quad (2.4.7)$$

then $g \in \Lambda_\alpha^*$. Let $x, y \in [-\pi, \pi]$, $x \neq y$, and we choose n so that $2^{-n-1} < |x - y| \leq 2^{-n}$. (There is no loss of generality in assuming that $|x - y| \leq 1/2$.) As before, we select a sequence of polynomials $S_k \in \mathbb{H}_{2^k}$ such that

$$\|g - S_k\|^* \leq 2E_{2^k}^*(g) \leq cB(g)2^{-k\alpha}, \quad k = 0, 1, 2, \dots$$

Then

$$\begin{aligned} |g(x) - g(y)| &\leq 2\|g - S_n\|^* + |S_n(x) - S_n(y)| \\ &\leq cB(g)2^{-n\alpha} + |x - y|\|S_n'\|^* \leq cB(g)2^{-n\alpha} + 2^{-n}\|S_n'\|^*. \end{aligned} \quad (2.4.8)$$

We may assume without loss of generality that $\|g\|^* \leq B(g)$. Now, using Bernstein inequality, we obtain (the reader should fill in the omitted details)

$$\begin{aligned}
\|S'_n\|^* &\leq \sum_{k=2}^n \|S'_k - S'_{k-1}\|^* + \|S'_1\|^* \\
&\leq \sum_{k=2}^n 2^k \|S_k - S_{k-1}\|^* + 2\|S_1\|^* \\
&\leq c \sum_{k=2}^n 2^{k(1-\alpha)} B(g) + c\|g\|^* \\
&\leq c2^{n(1-\alpha)} B(g).
\end{aligned}$$

Along with (2.4.8), this yields

$$|g(x) - g(y)| \leq cB(g)2^{-n\alpha} \leq cB(g)|x - y|^\alpha.$$

Thus, we have proved that (2.4.7) implies that $g \in \Lambda_\alpha^*$. Applying this fact with $f^{(r)}$ and using (2.4.6), we have completed the proof that (b) \Rightarrow (a). \square

3 Approximation with scattered data

3.1 Introduction

In Chapter 1, we studied the construction of the shifted average operator to achieve the rate of approximation which we observed to be optimal in Chapter 2. This operator was constructed using the Fourier coefficients of the target function. In this chapter, we will study the construction of a similar operator using values of the function. In classical signal processing and approximation theory, one typically chooses the “sites” at which the values are taken; e.g., at lattice points, or at the zeros of certain orthogonal polynomials, etc. In many modern applications, one does not have this liberty. Moreover, the values can be observed only approximately. Therefore, our aim is to obtain a stable method that depends upon observations of a function at an asymptotically optimal number of scattered sites. The immediate instinct is to discretize the integral expressions for the Fourier coefficients. However, in order to do this carefully so as not to disturb the degree of approximation and the number of sites, one needs powerful quadrature formulas. The main contribution of this chapter is to develop these formulas.

In Section 3.2, we prove some preliminary results in an abstract form. These will be applied in Section 3.3 in the setting of multivariate trigonometric polynomials to derive certain quadrature formulas. In Section 3.4, we will construct the desired approximation operators, known as quasi-interpolation operators.

3.2 Preliminary results

The results in this section and the next are motivated by the following quadrature formula:

$$\frac{1}{n+1} \sum_{\ell=0}^n T\left(\frac{2\pi\ell}{n+1}\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} T(t) dt, \quad T \in \mathbb{H}_n. \quad (3.2.1)$$

The formula (3.2.1) follows immediately from the following identities which can be verified easily for $k = 0, 1, \dots, \pm n$:

$$\frac{1}{n+1} \sum_{\ell=0}^n \exp\left(\frac{2\pi\ell ki}{n+1}\right) = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{if } k = \pm 1, \dots, \pm n. \end{cases}$$

In place of the univariate trigonometric polynomials, we wish to have multivariate trigonometric polynomials. More importantly, instead of the equidistant points $2\pi\ell/(n+1)$, we wish to have arbitrary points, and yet have weights with known bounds in place of $1/(n+1)$. We need some functional analysis to help us here.

Let $(X, \|\cdot\|_X)$ be a finite dimensional normed linear space, (e.g., $\mathbb{H}_{n,s}$) $(X^*, \|\cdot\|_{X^*})$ be its dual space, $\mathcal{Z} = \{x_1^*, \dots, x_M^*\} \subset X^* \setminus \{0\}$ (e.g., the point evaluation functionals at the scattered sites), and $x^* \in X^*$ (e.g., the functional that associates with every trigonometric polynomial its integral). In this section, we will study the question of finding nonnegative numbers w_1, \dots, w_M such that

$$x^*(x) = \sum_{\ell=1}^M w_\ell x_\ell^*(x), \quad x \in X.$$

Obviously, such a quadrature formula does not hold in general. In particular, if the quadrature formula holds, then x^* must be *positive with respect to \mathcal{Z}* ; i.e., if $x \in X$, and $x_\ell^*(x) \geq 0$ for $\ell = 1, \dots, M$, then $x^*(x) \geq 0$. It turns out that we need one more condition.

Definition 3.2.1 *Let*

$$S(x) := (x_1^*(x), \dots, x_M^*(x)), \quad x \in X, \quad (3.2.2)$$

$\|\cdot\|$ be a norm on \mathbb{R}^M and $\|\|\cdot\|\|^$ be its dual norm. The set \mathcal{Z} is called a norming set if there exists a constant $\alpha > 0$ such that*

$$\|x\|_X \leq \alpha \|\|\cdot\|\|^*(S(x)), \quad x \in X. \quad (3.2.3)$$

The main theorem of this section is the following.

Theorem 3.2.1 *Let \mathcal{Z} be a norming set. There exist real numbers w_1, \dots, w_M such that*

$$x^*(x) = \sum_{\ell=1}^M w_\ell x_\ell^*(x), \quad x \in X, \quad (3.2.4)$$

and

$$\|\|(w_1, \dots, w_M)\|\|^* \leq \alpha \|x^*\|_{X^*}. \quad (3.2.5)$$

If x^ is positive with respect to \mathcal{Z} , and there exists some $x_0 \in X$ such that $x_\ell^*(x_0) > 0$ for $\ell = 1, \dots, M$, then we may choose the w_ℓ 's in (3.2.4) nonnegative.*

We remark that the nonnegative weights might not satisfy (3.2.5). The proof of the second part of this theorem utilizes a consequence of the Hahn-Banach theorem known as the Krein-Rutman theorem (Theorem 3.2.2 below), which we state without proof. A vector space Y is called an ordered linear space if there is a relation $\preceq \subset Y \times Y$ with the following properties: \preceq is reflexive and transitive, and if $x \preceq y$ then $x + z \preceq y + z$ for all $z \in Y$, and $\alpha x \preceq \alpha y$ for all $\alpha \geq 0$. For example the space C_s^* is an ordered linear space with the ordering $f \preceq g$ if $f(x) \leq g(x)$ for all $x \in [-\pi, \pi]^s$. A functional ϕ on Y is called a positive functional if $x \preceq y$ implies $\phi(x) \leq \phi(y)$.

Theorem 3.2.2 *Let Y be an ordered linear space, M be a subspace of Y , $P := \{y \in Y : y \geq 0\}$, and $M \cap P$ have an interior point of P . Then any positive linear functional on M admits an extension as a positive linear functional on Y .*

PROOF OF THEOREM 3.2.1. The norming set property implies that S is injective. Let $V = S(X)$. Thus, we may define a functional $y : V \rightarrow \mathbb{R}$ by the formula $y(S(x)) := x^*(x)$, $x \in X$. We have

$$\begin{aligned} \sup_{v \in V, \|v\| \leq 1} |y(v)| &= \sup_{x \in X, \|S(x)\| \leq 1} |x^*(x)| \\ &\leq \sup_{x \in X, \|x\|_X \leq \alpha} |x^*(x)| \leq \alpha \|x^*\|_{X^*}. \end{aligned}$$

Therefore, by Hahn-Banach theorem, there exists an extension of the functional y to \mathbb{R}^M , which may be represented by (w_1, \dots, w_M) , such that $\|(w_1, \dots, w_M)\|^* \leq \alpha \|x^*\|_{X^*}$, and for $x \in X$,

$$x^*(x) = y(S(x)) = \sum_{\ell=1}^M w_\ell x_\ell^*(x).$$

For the positivity assertion, we use the Krein-Rutman theorem instead of the Hahn-Banach theorem. \square

REMARK One can always choose $\epsilon_\ell \in \{-1, 1\}$ such that for some $x_0 \in X$, $\epsilon_\ell x_\ell^*(x_0) > 0$ for $\ell = 1, \dots, M$. In order to prove this, we observe that if this were not the case then

$$X = \bigcup_{\ell=1}^M \{x \in X : x_\ell^*(x) = 0\}.$$

By the Baire category theorem, at least one of the sets $A_\ell := \{x \in X : x_\ell^*(x) = 0\}$ must have an interior point, z . Now, if $x \in X$ is arbitrary, then for a suitably small $\delta > 0$, $\delta(x - z) \in A_\ell$. This leads to $x_\ell^*(x) = 0$. Thus, the functional $x_\ell^* = 0$, a contradiction.

3.3 Quadrature formulas

Let $s \geq 1$ be a fixed integer, and \mathcal{C}_0 be a set of distinct points in $[-\pi, \pi]^s$, with the mesh norm defined by

$$\delta_{\mathcal{C}_0} := \max_{\mathbf{x} \in [-\pi, \pi]^s} \text{dist}(\mathbf{x}, \mathcal{C}_0) := \max_{\mathbf{x} \in [-\pi, \pi]^s} \min_{\mathbf{y} \in \mathcal{C}_0} |\mathbf{x} - \mathbf{y}|_\infty, \quad (3.3.1)$$

where

$$|\mathbf{x} - \mathbf{y}|_\infty := \max_{1 \leq j \leq s} |x_j - y_j|, \quad \mathbf{x} = (x_1, \dots, x_s), \quad \mathbf{y} = (y_1, \dots, y_s) \in \mathbb{R}^s.$$

The main result of this section is the following quadrature formula for $\mathbb{H}_{n,s}$ analogous to (3.2.1), where the equidistant nodes are replaced by points in \mathcal{C}_0 , and the equal weights are replaced by other suitable weights.

Theorem 3.3.1 *Let \mathcal{C}_0 be a set of distinct points in $[-\pi, \pi]^s$ and $n \geq 1$ be an integer such that $\delta_{\mathcal{C}_0} < \pi/(2 \cdot 3^{s+3}n)$. Then there exist numbers $\{w_\xi\}_{\xi \in \mathcal{C}_0}$ such that*

$$|w_\xi| \leq \frac{c}{n^s}, \quad \xi \in \mathcal{C}_0, \quad (3.3.2)$$

and for every $T \in \mathbb{H}_{n,s}$,

$$\frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} T(\mathbf{t}) d\mathbf{t} = \sum_{\xi \in \mathcal{C}_0} w_\xi T(\xi). \quad (3.3.3)$$

We may also choose w_ξ to be nonnegative instead of requiring (3.3.2).

In view of Theorem 3.2.1, this theorem will follow easily from the following Marcinkiewicz-Zygmund (-type) inequalities.

Theorem 3.3.2 *Let $\epsilon > 0$, and $n \geq 1$, $N \geq (4\pi \cdot 3^s + \epsilon)(n/\epsilon)$ be integers. Suppose \mathcal{C} is a set of points in $[-\pi, \pi]^s$ such that each cube with sides of length $2\pi/N$ contains exactly one point of \mathcal{C} . Let $\{R_\xi\}_{\xi \in \mathcal{C}}$ be a partition of $[-\pi, \pi]^s$, such that each R_ξ is a cube having side of length $2\pi/N$ and containing a point $\xi \in \mathcal{C}$. Then for any $T \in \mathbb{H}_{n,s}$,*

$$(1 - \epsilon) \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} |T(\xi)| \leq \int_{[-\pi, \pi]^s} |T(\mathbf{x})| d\mathbf{x} \leq (1 + \epsilon) \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} |T(\xi)|. \quad (3.3.4)$$

In order to prove Theorem 3.3.2, we first obtain certain technical estimates on the shifted average kernels V_n^* introduced in (1.3.3). In the remainder of this chapter, let $\mathbf{e} = (1, \dots, 1) \in \mathbb{Z}^s$, and

$$V_{n,s}^*(\mathbf{x}) = \prod_{j=1}^s V_n^*(x_j), \quad \mathbf{x} = (x_1, \dots, x_s). \quad (3.3.5)$$

Lemma 3.3.1 *Let $\eta \in (0, 1)$, $n, s \geq 1$, $N \geq n$ be integers, $\eta = n/N$. Suppose \mathcal{C} is a set of points in $[-\pi, \pi]^s$ such that each cube with sides of length $2\pi/N$ contains exactly one point of \mathcal{C} . Let $\{R_\xi\}_{\xi \in \mathcal{C}}$ be a partition of $[-\pi, \pi]^s$, such that each R_ξ is a cube having side of length $2\pi/N$ and containing a point $\xi \in \mathcal{C}$. Then*

$$\sup_{\mathbf{x} \in [-\pi, \pi]^s} \sum_{\xi \in \mathcal{C}} \int_{R_\xi} |V_{n,s}^*(\mathbf{u} - \mathbf{x}) - V_{n,s}^*(\xi - \mathbf{x})| d\mathbf{u} \leq 4\pi(6\pi)^s \frac{\eta(1 - \eta^s)}{1 - \eta}. \quad (3.3.6)$$

PROOF. Let $\mathbf{x} \in [-\pi, \pi]^s$. Since we may take the partition $\{R_\xi + \mathbf{x}\}$ in place of $\{R_\xi\}$, there is no loss of generality in assuming that $\mathbf{x} = \mathbf{0}$. Let $\mathbf{c}_\xi = (c_{\xi,1}, \dots, c_{\xi,s})$ be the center of R_ξ , $\xi \in \mathcal{C}$. For each $\xi \in \mathcal{C}$, we have

$$\begin{aligned}
& \int_{R_\xi} |V_{n,s}^*(\mathbf{u}) - V_{n,s}^*(\xi)| d\mathbf{u} \\
& \leq \int_{R_\xi} \left| \sum_{\ell=1}^s \prod_{j=1}^{\ell-1} V_n^*(\xi_j) \prod_{j=\ell+1}^s V_n^*(u_j) (V_n^*(u_\ell) - V_n^*(\xi_\ell)) \right| d\mathbf{u} \\
& \leq \sum_{\ell=1}^s \left(\prod_{j=1}^{\ell-1} \left(\frac{2\pi}{N} |V_n^*(\xi_j)| \right) \prod_{j=\ell+1}^s \int_{c_{\xi,j}-\pi/N}^{c_{\xi,j}+\pi/N} |V_n^*(u_j)| du_j \right. \\
& \quad \left. \times \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} |V_n^*(u_\ell) - V_n^*(\xi_\ell)| du_\ell \right). \tag{3.3.7}
\end{aligned}$$

Now, in view of (1.3.6), we see that

$$\frac{2\pi}{N} |V_n^*(\xi_j)| \leq \frac{6\pi n}{N}, \quad j = 1, \dots, s,$$

and

$$\begin{aligned}
& \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} |V_n^*(u_\ell) - V_n^*(\xi_\ell)| du_\ell \\
& = \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} \left| \int_{\xi_\ell}^{u_\ell} V_n^{*\prime}(t) dt \right| du_\ell \\
& \leq \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} |V_n^{*\prime}(t)| dt du_\ell \\
& = \frac{2\pi}{N} \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} |V_n^{*\prime}(t)| dt
\end{aligned}$$

Therefore, (3.3.7) implies that

$$\begin{aligned}
& \int_{R_\xi} |V_{n,s}^*(\mathbf{u}) - V_{n,s}^*(\xi)| d\mathbf{u} \\
& \leq \frac{2\pi}{N} \sum_{\ell=1}^s \left(\frac{6\pi n}{N} \right)^{\ell-1} \prod_{j=\ell+1}^s \int_{c_{\xi,j}-\pi/N}^{c_{\xi,j}+\pi/N} |V_n^*(u_j)| du_j \int_{c_{\xi,\ell}-\pi/N}^{c_{\xi,\ell}+\pi/N} |V_n^{*\prime}(t)| dt.
\end{aligned}$$

Hence, using (1.3.6) and Bernstein's inequality (2.2.2) with $p = 1$, we deduce that

$$\begin{aligned}
& \sum_{\xi \in \mathcal{C}} \int_{R_\xi} |V_{n,s}^*(\mathbf{u}) - V_{n,s}^*(\xi)| d\mathbf{u} \\
& \leq \frac{2\pi}{N} \sum_{\ell=1}^s \left(\frac{6\pi n}{N} \right)^{\ell-1} \prod_{j=\ell+1}^s \int_{-\pi}^{\pi} |V_n^*(u_j)| du_j \int_{-\pi}^{\pi} |V_n^{*\prime}(t)| dt
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{4\pi n}{N} \sum_{\ell=1}^s (6\pi)^{s-1} \left(\frac{n}{N}\right)^{\ell-1} \int_{-\pi}^{\pi} |V_n^*(t)| dt \\
&\leq 4\pi (6\pi)^s \frac{\eta(1-\eta^s)}{1-\eta}.
\end{aligned}$$

□

PROOF OF THEOREM 3.3.2. We observe that $\eta := n/N \leq \epsilon/(4\pi \cdot 3^s + \epsilon)$. Let $T \in \mathbb{H}_{n,s}$. By a repeated application of Theorem 1.3.1 to each coordinate, we obtain the representation

$$T(\mathbf{u}) = \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} T(\mathbf{x}) V_{n,s}^*(\mathbf{u} - \mathbf{x}) d\mathbf{x}, \quad \mathbf{u} \in [-\pi, \pi]^s. \quad (3.3.8)$$

Hence, using Lemma 3.3.1, we deduce that

$$\begin{aligned}
&\left| \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} |T(\xi)| - \int_{[-\pi, \pi]^s} |T(\mathbf{u})| d\mathbf{u} \right| \\
&\leq \sum_{\xi \in \mathcal{C}} \int_{R_\xi} |T(\xi) - T(\mathbf{u})| d\mathbf{u} \\
&= \frac{1}{(2\pi)^s} \sum_{\xi \in \mathcal{C}} \int_{R_\xi} \left| \int_{[-\pi, \pi]^s} T(\mathbf{x}) V_{n,s}^*(\xi - \mathbf{x}) d\mathbf{x} - \int_{[-\pi, \pi]^s} T(\mathbf{x}) V_{n,s}^*(\mathbf{u} - \mathbf{x}) d\mathbf{x} \right| d\mathbf{u} \\
&\leq \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} |T(\mathbf{x})| \sum_{\xi \in \mathcal{C}} \int_{R_\xi} |V_{n,s}^*(\xi - \mathbf{x}) - V_{n,s}^*(\mathbf{u} - \mathbf{x})| d\mathbf{u} d\mathbf{x} \\
&\leq 4\pi \cdot 3^s \frac{\eta(1-\eta^s)}{1-\eta} \int_{[-\pi, \pi]^s} |T(\mathbf{x})| d\mathbf{x}. \quad (3.3.9)
\end{aligned}$$

In view of our choice of η , $\eta/(1-\eta) \leq \epsilon/(4\pi 3^s)$. Hence, (3.3.9) leads to (3.3.4). □

PROOF OF THEOREM 3.3.1. In this proof, we let $\epsilon = 1/2$, and $N := 3^{s+3}n$. Then $(4\pi \cdot 3^s + \epsilon)/\epsilon < 8\pi \cdot 3^s + 1 < 3^{s+3}$, and $N \geq (4\pi \cdot 3^s + \epsilon)(n/\epsilon)$. Moreover, $2\delta_{\mathcal{C}_0} \leq \pi/N$. We divide $[-\pi, \pi]^s$ into congruent subcubes of side $2\pi/N$. Each of these subcubes has at least one point of \mathcal{C}_0 , and we may choose a subset \mathcal{C} , so that each subcube has exactly one point of \mathcal{C}_0 . It is easy to verify that $\delta_{\mathcal{C}_0} \leq \delta_{\mathcal{C}} \leq 2\delta_{\mathcal{C}_0} \leq \pi/N$. Thus, all the hypotheses of Theorem 3.3.2 are satisfied, and (3.3.4) holds.

Now, we apply Theorem 3.2.1 with the following choices. We let X be the space $\mathbb{H}_{n,s}$, with

$$\|T\|_X = \int_{[-\pi, \pi]^s} |T(\mathbf{t})| d\mathbf{t}.$$

For each $\xi \in \mathcal{C}$, we define $x_\xi^*(T) := T(\xi)$, $T \in \mathbb{H}_{n,s}$, and $Z := \{x_\xi^*\}_{\xi \in \mathcal{C}}$. Further, we let

$$x^*(T) := \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} T(\mathbf{x}) d\mathbf{x},$$

and observe that $\|x^*\|_{X^*} = (2\pi)^{-s}$. Finally, let $M := |\mathcal{C}|$, and

$$\|(y_1, \dots, y_M)\| := \sum_{k=1}^M |y_k|.$$

Then

$$\|(y_1, \dots, y_M)\|^* := \max_{1 \leq k \leq M} |y_k|. \quad (3.3.10)$$

With these choices, (3.3.4) implies that Z is a norming set with $\alpha := (3/2) \left(\frac{2\pi}{N}\right)^s$. The polynomial $T_0(\mathbf{x}) = 1$ is in $\mathbb{H}_{n,s}$ and satisfies $x_\ell^*(T_0) = 1 > 0$ for $\ell = 1, \dots, M$. Finally, we verify that x^* is positive with respect to Z . Let $T \in \mathbb{H}_{n,s}$ and $T(\xi) \geq 0$ for all $\xi \in \mathcal{C}$. Arguing as in (3.3.9), and using (3.3.4), we see that

$$\begin{aligned} & \left| \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} T(\xi) - \int_{[-\pi, \pi]^s} T(\mathbf{u}) d\mathbf{u} \right| \\ & \leq 4\pi \cdot 3^s \frac{\eta(1 - \eta^s)}{1 - \eta} \int_{[-\pi, \pi]^s} |T(\mathbf{x})| d\mathbf{x} \\ & \leq \epsilon(1 + \epsilon) \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} T(\xi) \\ & \leq (3/4) \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} T(\xi). \end{aligned}$$

Hence,

$$\int_{[-\pi, \pi]^s} T(\mathbf{u}) d\mathbf{u} \geq (1/4) \left(\frac{2\pi}{N}\right)^s \sum_{\xi \in \mathcal{C}} T(\xi) \geq 0.$$

This proves that x^* is positive with respect to Z . Thus, all the hypotheses of Theorem 3.2.1 are satisfied. This implies the existence of nonnegative numbers w_ξ , $\xi \in \mathcal{C}$, such that equation (3.2.4) implies (3.3.3). If the w_ξ 's are not required to be nonnegative, the estimate (3.2.5) takes the form (3.3.2). We define $w_\xi := 0$ if $\xi \in \mathcal{C}_0 \setminus \mathcal{C}$. \square

3.4 Quasi-interpolatory operators

In this section, we discretize the shifted average operators to obtain their analogues that depend upon the values of the target function at scattered sites.

Theorem 3.4.1 *Let \mathcal{C}_0 be a set of distinct points in $[-\pi, \pi]^s$ and $n \geq 1$ be an integer such that $\delta_{\mathcal{C}_0} < \pi/(2 \cdot 3^{s+4}n)$.*

(a) *There exist numbers $\{w_\xi\}_{\xi \in \mathcal{C}_0}$ such that*

$$|w_\xi| \leq cn^{-s}, \quad \xi \in \mathcal{C}_0, \quad (3.4.1)$$

and

$$\frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} T(\mathbf{t}) d\mathbf{t} = \sum_{\xi \in \mathcal{C}_0} w_\xi T(\xi), \quad T \in \mathbb{H}_{3n, s}. \quad (3.4.2)$$

(b) Let the operator $v_{n, s}^*$ be defined by

$$v_{n, s}^*(f, \mathbf{x}) := \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} f(\mathbf{t}) V_{n, s}^*(\mathbf{x} - \mathbf{t}) d\mathbf{t}. \quad (3.4.3)$$

Then $v_{n, s}^*(T) = T$ for every $T \in \mathbb{H}_{n, s}$. Also, for $f \in C_s^*$, $v_{n, s}^*(f) \in \mathbb{H}_{2n-1, s}$, and we have

$$\|v_{n, s}^*(f)\|_s^* \leq c \|f\|_s^*, \quad E_{2n-1, s}^*(f) \leq \|f - v_{n, s}^*(f)\|_s^* \leq c E_{n, s}^*(f). \quad (3.4.4)$$

(c) Let the operator $\tau_{n, s}^*$ be defined by

$$\tau_{n, s}^*(f, \mathbf{x}) := \tau_{n, s}^*(\mathcal{C}_0; f, \mathbf{x}) := \sum_{\xi \in \mathcal{C}_0} w_\xi f(\xi) V_{n, s}^*(\mathbf{x} - \xi), \quad f \in C_s^*. \quad (3.4.5)$$

Then $\tau_{n, s}^*(T) = T$ for every $T \in \mathbb{H}_{n, s}$. Also, for $f \in C_s^*$, $\tau_{n, s}^*(f) \in \mathbb{H}_{2n-1, s}$, and we have

$$\|\tau_{n, s}^*(f)\|_s^* \leq c \|f\|_s^*, \quad E_{2n-1, s}^*(f) \leq \|f - \tau_{n, s}^*(f)\|_s^* \leq c E_{n, s}^*(f). \quad (3.4.6)$$

PROOF. Part (a) is simply Theorem 3.3.1 applied with $3n$ in place of n . We note further that Theorem 3.3.2 also applies with $3n$ in place of n . The proof of part (b) is similar to that of Theorem 1.3.1. We omit the details. Let $T \in \mathbb{H}_{n, s}$. Using (3.4.2) with the polynomial $T(\mathbf{t})V_{n, s}^*(\mathbf{x} - \mathbf{t})$ of order $3n$ in \mathbf{t} , we get

$$T(\mathbf{x}) = \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} T(\mathbf{t}) V_{n, s}^*(\mathbf{x} - \mathbf{t}) d\mathbf{t} = \sum_{\xi \in \mathcal{C}_0} w_\xi T(\xi) V_{n, s}^*(\mathbf{x} - \xi) = \tau_{n, s}^*(T, \mathbf{x}). \quad (3.4.7)$$

It is clear that $\tau_{n, s}^*(f) \in \mathbb{H}_{2n-1, s}$ for every $f \in C_s^*$. The estimates (3.4.1) and Theorem 3.3.2 imply that

$$\begin{aligned} \sum_{\xi \in \mathcal{C}_0} |w_\xi V_{n, s}^*(\mathbf{x} - \xi)| &\leq \frac{c}{(2n)^s} \sum_{\xi \in \mathcal{C}_0} |V_{n, s}^*(\mathbf{x} - \xi)| \\ &\leq c \int_{[-\pi, \pi]^s} |V_{n, s}^*(\mathbf{x} - \mathbf{t})| d\mathbf{t} \leq c. \end{aligned}$$

The first estimate in (3.4.6) now follows from the definition of $\tau_{n, s}^*$. The other estimates follow from this and (3.4.7) exactly as in the proof of Theorem 1.3.1. We omit the details again. \square

We point out one way to obtain the quantities w_ξ numerically. We solve the linear programming problem:

$$\begin{aligned} \text{Minimize } \max_{\xi \in \mathcal{C}_0} |w_\xi| \text{ subject to the conditions (3.4.2), with } e^{i\mathbf{k} \cdot \mathbf{x}} \text{ in place of } T, \\ \mathbf{k} \in \{0, \pm 1, \dots, \pm(3n)\}^s. \end{aligned}$$

Theorem 3.3.1 shows that this problem has a feasible solution satisfying (3.4.1).

4 Approximation with periodic neural networks

4.1 Introduction

A neural network is a device for a highly parallel computation of functions. A basic ingredient of a neural network is a *neuron*. A neuron is a special purpose computer with a local memory that can take any finite number of inputs, and evaluate a special function based on the inputs and the content of its memory. Typically, when an input from \mathbb{R}^s is expected, one stores certain *weights* $\mathbf{w} \in \mathbb{R}^s$ and a *threshold* $b \in \mathbb{R}$ in the memory. Upon receiving the input $\mathbf{x} \in \mathbb{R}^s$, the neuron computes the inner product $\mathbf{w} \cdot \mathbf{x}$, and evaluates a certain *activation function* $\phi : \mathbb{R} \rightarrow \mathbb{R}$ to obtain $\phi(\mathbf{w} \cdot \mathbf{x} + b)$. This expression is the output of the neuron. In a neural network, one arranges a number of neurons in whatever *network topology*. Thus, we may think of a neural network as a directed graph with neurons as its nodes. The edge from one neuron n_1 to another, n_2 , in this “topology” indicates that the output of n_1 is one of the inputs of n_2 . Special I/O devices furnish input to the network and presents an output of the network to the user.

A simple, and most commonly used, network topology is known as a *feedforward network with a single hidden layer* (cf. Figure 2). Assuming N neurons in the hidden

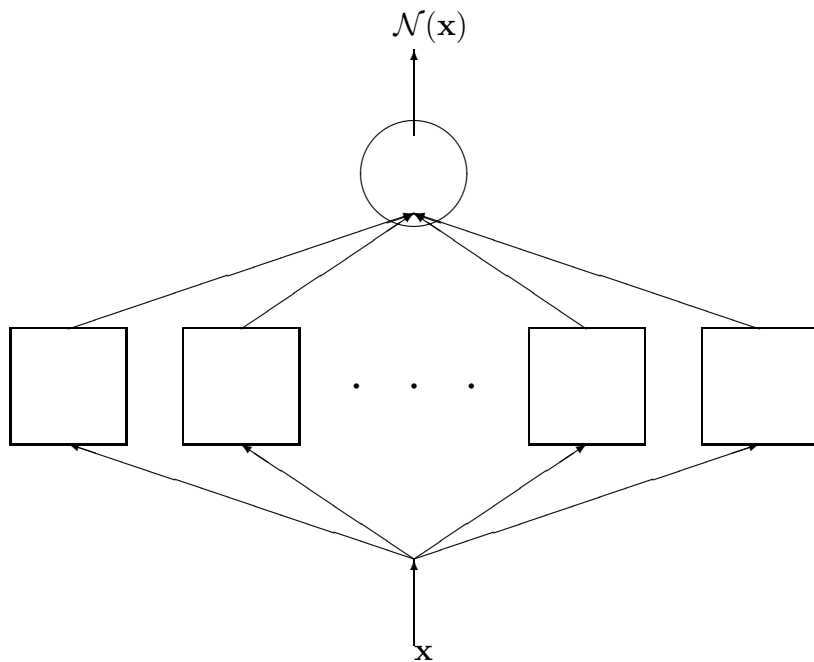


Figure 2: A feedforward neural network with one hidden layer. The rectangular nodes evaluate a nonlinearity of the form $\phi(\mathbf{w} \cdot \mathbf{x} + b)$, the circular node evaluates a linear combination of its inputs.

layer, the output of the network upon an input of $\mathbf{x} \in \mathbb{R}^s$ is of the form $\sum_{k=1}^N a_k \phi(\mathbf{w}_k \cdot$

$\mathbf{x} + b_k$). The set of all functions $\mathbf{x} \mapsto \sum_{k=1}^N a_k \phi(\mathbf{w}_k \cdot \mathbf{x} + b_k)$, with $a_k, b_k \in \mathbb{R}$, $\mathbf{w}_k \in \mathbb{R}^s$, $1 \leq k \leq N$, will be denoted by $\Pi_{\phi;N,s}$. Some of the most popular activation functions are: *the squashing function* : $\phi(x) := (1 + e^{-x})^{-1}$, *the hard threshold function*: $\phi(x) := 1$, if $x \geq 0$, and 0 otherwise, and the *linear threshold function*, which is equal to the hard threshold function outside $[-1, 1]$, and linear on this interval.

A process by which the weights, thresholds, and coefficients are selected is known as a *training (or learning) method*. The most widely used learning method for a function using its values at scattered sites (the *training data*) is to do a least squared fit. One reason for the popularity of choosing ϕ to be the squashing function is that its derivative can be easily expressed in terms of the function itself: $\phi' = \phi(1 - \phi)$. Therefore, the nonlinear equations involved in the least squared fit become computationally easier to solve using such methods as the steepest descent method. In the neural network literature, this is referred to as the *back-propagation method*. Some of the disadvantages of this method are the following. The network topology and the size of the network has to be fixed in advance in order to set up the equations. There is no guarantee that the method will not find a local minimum rather than the global minimum of the least squared error. There is no guarantee that the network will *generalize* properly; i.e., work well to predict the value of the function outside the training data.

Neural networks are ubiquitous in military and civilian applications including robotics, image processing, speech recognition and reproduction, automatic control of airplanes and missiles, automated target recognition and tracking, etc. The most often quoted reason for this popularity is their ability to approximate “arbitrary” functions. All the classical questions of approximation theory are thus interesting in the context of approximation by elements of $\Pi_{\phi;N,s}$. In this chapter, we will explore the case of approximation of periodic functions using periodic networks; i.e., networks with a periodic activation function, where the weights are restricted to be integers. As before, we are mostly interested in the complexity problem, and the problem of constructing good approximation.

In Section 4.2, we establish a close connection between trigonometric approximation and approximation by periodic networks. Along with the results of Section 3.4, we will give a training method for networks without using any kind of nonlinear optimization involving the activation function. In particular, our method has none of shortfalls of these optimization based methods. In Section 4.3, we point out the analogues and extensions in the case of approximation on the sphere.

4.2 Approximation by periodic networks

Let $\phi \in C^*$, and

$$\Pi_{\phi;N,s}^* := \left\{ \sum_{k=1}^N a_k \phi(\mathbf{w}_k \cdot (\cdot) + b_k) : a_k, b_k \in \mathbb{R}, \mathbf{w}_k \in \mathbb{Z}^s, 1 \leq k \leq N \right\}. \quad (4.2.1)$$

The basis of our results in this section is the following fundamental proposition.

Proposition 4.2.1 *Let $\phi \in C^*$ and $c_1^*(\phi) \neq 0$. Then for any integer $N \geq 1$,*

$$\left\| e^{i\cdot} - \frac{1}{(2N+1)c_1^*(\phi)} \sum_{k=0}^{2N} \exp\left(\frac{2ik\pi}{2N+1}\right) \phi\left(\cdot - \frac{2\pi k}{2N+1}\right) \right\|^* \leq \frac{4}{|c_1^*(\phi)|} E_N^*(\phi). \quad (4.2.2)$$

PROOF. From the definition of $c_1^*(\phi)$, we have for $x \in [-\pi, \pi]$,

$$e^{ix} = \frac{1}{2\pi c_1^*(\phi)} \int_{-\pi}^{\pi} \phi(t) e^{i(x-t)} dt = \frac{1}{2\pi c_1^*(\phi)} \int_{-\pi}^{\pi} \phi(x-t) e^{it} dt. \quad (4.2.3)$$

Now, the key observation is that for any $N \geq 1$,

$$\int_{-\pi}^{\pi} \phi(x-t) e^{it} dt = \int_{-\pi}^{\pi} v_N^*(\phi, x-t) e^{it} dt.$$

As a function of t , $v_N^*(\phi, x-t) e^{it} \in \mathbb{H}_{2N}$. So, we may evaluate the last integral using the quadrature formula (3.2.1) to obtain

$$c_1^*(\phi) e^{ix} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(x-t) e^{it} dt = \frac{1}{(2N+1)} \sum_{k=0}^{2N} \exp\left(\frac{2ik\pi}{2N+1}\right) v_N^*\left(\phi, \left(x - \frac{2\pi k}{2N+1}\right)\right).$$

Now, using (1.3.5), we obtain for all $x \in [-\pi, \pi]$:

$$\begin{aligned} & \left| \frac{1}{(2N+1)} \sum_{k=0}^{2N} \exp\left(\frac{2ik\pi}{2N+1}\right) v_N^*\left(\phi, \left(x - \frac{2\pi k}{2N+1}\right)\right) \right. \\ & \quad \left. - \frac{1}{(2N+1)} \sum_{k=0}^{2N} \exp\left(\frac{2ik\pi}{2N+1}\right) \phi\left(x - \frac{2\pi k}{2N+1}\right) \right| \\ & \leq 4E_N^*(\phi). \end{aligned}$$

Along with the previous equation, this leads to (4.2.2). \square

Using Proposition 4.2.1, we may obtain a neural network to approximate any trigonometric polynomial. We define for $f \in C_s^*$ and $\mathbf{j} \in \mathbb{Z}^s$,

$$c_{\mathbf{j}}^*(f) := \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} f(\mathbf{x}) e^{-i\mathbf{j} \cdot \mathbf{x}} d\mathbf{x}, \quad (4.2.4)$$

and for integers $N, n \geq 1$,

$$\mathcal{N}_{N,n,s}(\phi; f, \mathbf{x}) := \frac{1}{(2N+1)c_1^*(\phi)} \sum_{k=0}^{2N} \sum_{-n \leq \mathbf{j} \leq n} c_{\mathbf{j}}^*(f) \exp\left(\frac{2ik\pi}{2N+1}\right) \phi\left(\mathbf{j} \cdot \mathbf{x} - \frac{2\pi k}{2N+1}\right). \quad (4.2.5)$$

The function $\mathcal{N}_{N,n,s}(\phi; f) \in \Pi_{\phi; (2N+1)(2n+1)^s, s}$.

Theorem 4.2.1 *Let $s, n, N \geq 1$ be integers, and $T \in \mathbb{H}_{n,s}$. Then*

$$\|T - \mathcal{N}_{N,n,s}(\phi; T)\|_s^* \leq \frac{4(2n+1)^{s/2} E_N^*(\phi)}{|c_1^*(\phi)|} \|T\|_s^*. \quad (4.2.6)$$

PROOF. In this proof only, we write $e_{\mathbf{k}}(\mathbf{x}) := \exp(i\mathbf{k} \cdot \mathbf{x})$, $\mathbf{k} \in \mathbb{Z}^s$. Proposition 4.2.1 implies that for $-n \leq \mathbf{k} \leq n$,

$$\|e_{\mathbf{k}} - \mathcal{N}_{N,n,s}(\phi; e_{\mathbf{k}})\|_s^* \leq \frac{4E_N^*(\phi)}{|c_1^*(\phi)|}. \quad (4.2.7)$$

We observe that

$$\mathcal{N}_{N,n,s}(\phi; T) = \sum_{-n \leq \mathbf{k} \leq n} c_{\mathbf{k}}^*(T) \mathcal{N}_{N,n,s}(\phi; e_{\mathbf{k}}),$$

and hence, (4.2.7) implies that

$$\|T - \mathcal{N}_{N,n,s}(\phi; T)\|_s^* \leq \frac{4E_N^*(\phi)}{|c_1^*(\phi)|} \sum_{-n \leq \mathbf{k} \leq n} |c_{\mathbf{k}}^*(T)|. \quad (4.2.8)$$

Now, we recall the Parseval identity, which states that

$$\sum_{-n \leq \mathbf{k} \leq n} |c_{\mathbf{k}}^*(T)|^2 = \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} |T(\mathbf{x})|^2 d\mathbf{x}.$$

So, using Schwarz inequality, we obtain

$$\begin{aligned} \sum_{-n \leq \mathbf{k} \leq n} |c_{\mathbf{k}}^*(T)| &\leq (2n+1)^{s/2} \left\{ \sum_{-n \leq \mathbf{k} \leq n} |c_{\mathbf{k}}^*(T)|^2 \right\}^{1/2} \\ &= (2n+1)^{s/2} \left\{ \frac{1}{(2\pi)^s} \int_{[-\pi, \pi]^s} |T(\mathbf{x})|^2 d\mathbf{x} \right\}^{1/2} \\ &\leq (2n+1)^{s/2} \|T\|_s^*. \end{aligned}$$

□

The following theorem gives a training method for periodic functions, as well as assesses its *generalization capability*, i.e., the degree of approximation provided by the network at sites other than the those where the training data is collected.

Theorem 4.2.2 *Let $s \geq 1$ be an integer, \mathcal{C}_0 be a set of distinct points in $[-\pi, \pi]^s$ and $n \geq 1$ be an integer such that $\delta_{\mathcal{C}_0} < \pi/(2 \cdot 3^{s+4n})$. Let $\phi \in C^*$ and $c_1^*(\phi) \neq 0$. Then for any $f \in C^*$ and integer $N \geq 1$, we have*

$$\|f - \mathcal{N}_{N,2n-1,s}(\phi; \tau_{n,s}^*(f))\|_s^* \leq c \left\{ E_{n,s}^*(f) + \frac{n^{s/2} E_N^*(\phi)}{|c_1^*(\phi)|} \|f\|_s^* \right\}, \quad (4.2.9)$$

where $\tau_{n,s}^*(f)$ is defined as in Theorem 3.4.1.

PROOF. Using Theorem 3.4.1 and Theorem 4.2.1, we see that

$$\begin{aligned} \|f - \mathcal{N}_{N,2n-1,s}(\phi; \tau_{n,s}^*(f))\|_s^* &\leq \|f - \tau_{n,s}^*(f)\|_s^* + \|\tau_{n,s}^*(f) - \mathcal{N}_{N,2n-1,s}(\phi; \tau_{n,s}^*(f))\|_s^* \\ &\leq c E_{n,s}^*(f) + \frac{4(4n-1)^{s/2} E_N^*(\phi)}{|c_1^*(\phi)|} \|\tau_{n,s}^*(f)\|_s^* \\ &\leq c \left\{ E_{n,s}^*(f) + \frac{n^{s/2} E_N^*(\phi)}{|c_1^*(\phi)|} \|f\|_s^* \right\}. \end{aligned}$$

□

We illustrate the above theorem with an example. Let $\sigma(x) := (1 + e^{-x})^{-1}$, $\lambda(x) := \sigma(x+1) - \sigma(x-1)$. Then λ is integrable. The Fourier transform of λ can be computed using contour integration, which shows that

$$\hat{\lambda}(1) := \lim_{R \rightarrow \infty} \frac{1}{2\pi} \int_{-R}^R \lambda(x) e^{-ix} dx \neq 0.$$

We construct a periodization of λ by

$$\phi^{[sq]}(x) := \sum_{k \in \mathbb{Z}} \lambda(x - 2\pi k). \quad (4.2.10)$$

Since $|\lambda(x)| \leq (e - e^{-1})e^{-|x|}$ for $x \in \mathbb{R}$, the series in (4.2.10) converges uniformly on compact subsets of \mathbb{R} . The function $\phi^{[sq]}$ is clearly 2π -periodic, and one can compute easily that $c_1^*(\phi^{[sq]}) = \hat{\lambda}(1) \neq 0$. Further, it can be shown that there exists $\alpha > 0$ such that $E_N^*(\phi^{[sq]}) \leq e^{-\alpha N}$, $N = 1, 2, \dots$.

Now let $f : [-1, 1]^s \rightarrow \mathbb{R}$ be r times continuously differentiable. According to Whitney extension theorem, there exists an extension of f , $g : [-4, 4]^s \rightarrow \mathbb{R}$, such that

$$\sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-4, 4]^s} |D^j g(\mathbf{x})| \leq c \sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-1, 1]^s} |D^j f(\mathbf{x})|.$$

We now let ψ be an infinitely many times continuously differentiable function such that $\psi(\mathbf{x}) = 1$ if $\mathbf{x} \in [-1, 1]^s$ and equal to 0 outside $[-\pi/2, \pi/2]^s$. Then the function ψg has the properties that $\psi(\mathbf{x})g(\mathbf{x}) = f(\mathbf{x})$ for $\mathbf{x} \in [-1, 1]^s$, and

$$\sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-\pi, \pi]^s} |D^j(\psi g)(\mathbf{x})| \leq c \sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-1, 1]^s} |D^j f(\mathbf{x})|.$$

Further, since $\psi(\mathbf{x})g(\mathbf{x}) = 0$ outside $[-\pi/2, \pi/2]^s$, we may extend ψg as function on \mathbb{R}^s that is 2π -periodic in each of its variables. Denoting this extension by f^* , we see that $f^*(\mathbf{x}) = f(\mathbf{x})$ for $\mathbf{x} \in [-1, 1]^s$, and

$$\sum_{j=1}^s \|D_j^r f^*\|_s^* \leq c \sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-1, 1]^s} |D^j f(\mathbf{x})|. \quad (4.2.11)$$

Using Theorem 4.2.2 and Theorem 1.4.1, and taking $N := \frac{(r + s/2)}{\alpha} \log n$, we obtain that

$$\begin{aligned} & \max_{\mathbf{x} \in [-1, 1]^s} |f(\mathbf{x}) - \mathcal{N}_{N, 2n-1, s}(\phi^{[sq]}; \tau_{n, s}^*(f^*), \mathbf{x})| \\ & \leq \|f^* - \mathcal{N}_{N, 2n-1, s}(\phi^{[sq]}; \tau_{n, s}^*(f^*))\|_s^* \\ & \leq c \left\{ E_{n, s}^*(f^*) + \frac{n^{s/2} E_N^*(\phi^{[sq]})}{|c_1^*(\phi^{[sq]})|} \|f^*\|_s^* \right\} \\ & \leq c (n^{-r} + n^{s/2} e^{-\alpha N}) \sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-1, 1]^s} |D^j f(\mathbf{x})| \\ & \leq cn^{-r} \sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-1, 1]^s} |D^j f(\mathbf{x})|. \end{aligned} \quad (4.2.12)$$

Next, we observe that

$$|\phi^{[sq]}(x) - \sum_{|k| \leq M} \lambda(x - 2\pi k)| \leq ce^{-2\pi(M-|x|)}, \quad x \in \mathbb{R}.$$

If we choose $M = 2n$ and replace each occurrence of $\phi^{[sq]}(\mathbf{j} \cdot \mathbf{x} - (2\pi k)/(2N + 1))$ in $\mathcal{N}_{N,2n-1,s}(\phi^{[sq]}; \tau_{n,s}^*(f^*), \mathbf{x})$ by its partial sum, we obtain a network $G(f)$ having $cn^{s+1} \log n$ neurons. Imitating the proof of Theorem 4.2.1, we can prove that

$$\max_{\mathbf{x} \in [-1,1]^s} |\mathcal{N}_{N,2n-1,s}(\phi^{[sq]}; \tau_{n,s}^*(f^*), \mathbf{x}) - G(f, \mathbf{x})| \leq cn^{s/2} e^{-c_1 n} \leq c_2 e^{-c_3 n}.$$

Thus, (4.2.12) leads to a network $G(f)$ with $\mathcal{O}(n^{s+1} \log n)$ neurons such that

$$\max_{\mathbf{x} \in [-1,1]^s} |f(\mathbf{x}) - G(f, \mathbf{x})| \leq cn^{-r} \sum_{0 \leq j \leq r} \max_{\mathbf{x} \in [-1,1]^s} |D^{\mathbf{j}} f(\mathbf{x})|. \quad (4.2.13)$$

We observe that most of the complicated constructions involving partial sums etc. are done independently of the function. In fact, we have constructed $\mathcal{O}(n^s)$ basic networks, each containing $\mathcal{O}(n \log n)$ neurons such that for every data, the network to approximate the underlying function is simply a linear combination of these basic networks, with the values of the function as the coefficients.

We remark also that our construction involve only the minimal assumptions on the activation function, without which the classes $\Pi_{\phi;n,s}^*$ are not dense in C_s^* . Similar constructions can be made also for radial basis function networks, where the results are somewhat more impressive.

4.3 Approximation on the sphere

The ideas in Section 4.2 turn out to be much more fruitful in the context of approximation on the sphere. Let $q \geq 1$ be an integer, \mathbb{S}^q be the unit sphere in \mathbb{R}^{q+1} . In this section, all constants will depend upon q as well. Let $\phi : [-1, 1] \rightarrow \mathbb{R}$. A *zonal function network* (ZF network) is defined to be a finite linear combination of functions of the form $\mathbf{x} \mapsto \phi(\mathbf{x} \cdot \mathbf{y})$, $\mathbf{y} \in \mathbb{S}^q$. Our goal in this section is to construct ZF networks to approximate a real valued function on \mathbb{S}^q using its values at scattered sites.

First, we recall some facts. We denote the volume (surface area) measure on \mathbb{S}^q by μ_q , and the volume of \mathbb{S}^q by ω_q . We have

$$\omega_q := \int_{\mathbb{S}^q} d\mu_q = \frac{2\pi^{(q+1)/2}}{\Gamma((q+1)/2)}.$$

We introduce some classes of functions. For measurable $f : \mathbb{S}^q \rightarrow \mathbb{R}$, and $1 \leq p \leq \infty$, let

$$\|f\|_{\mathbb{S}^q, p} := \begin{cases} \left\{ \int_{\mathbb{S}^q} |f(\mathbf{x})|^p d\mu_q(\mathbf{x}) \right\}^{1/p}, & \text{if } 1 \leq p < \infty, \\ \text{ess sup}_{\mathbf{x} \in \mathbb{S}} |f(\mathbf{x})|, & \text{if } p = \infty. \end{cases}$$

The class of all measurable functions $f : \mathbb{S}^q \rightarrow \mathbb{C}$ for which $\|f\|_{\mathbb{S}^q, p} < \infty$ will be denoted by $L^p(\mathbb{S}^q)$, with the usual understanding that functions that are equal almost everywhere are considered equal as elements of $L^p(\mathbb{S}^q)$. All continuous complex valued functions on \mathbb{S}^q will be denoted by $C(\mathbb{S}^q)$.

We will also need the following norms and the corresponding function spaces on $[-1, 1]$, with weight function $w_q(x) := (1 - x^2)^{\frac{q}{2}-1}$:

$$\langle f, g \rangle_{w_q} := \int_{-1}^1 f(x) \overline{g(x)} w_q(x) dx, \quad (4.3.1)$$

$$\|f\|_{w_q, p} := \begin{cases} \left\{ \int_{-1}^1 |f(x)|^p w_q(x) dx \right\}^{1/p}, & \text{if } 1 \leq p < \infty, \\ \text{ess sup}_{x \in [-1, 1]} |f(x)|, & \text{if } p = \infty. \end{cases} \quad (4.3.2)$$

A homogeneous, harmonic polynomial of degree ℓ is called a spherical harmonic. Let \mathbf{H}_ℓ^q be the class of all spherical harmonics of degree ℓ . The dimension d_ℓ^q of \mathbf{H}_ℓ^q is given by

$$d_\ell^q = \begin{cases} \frac{2\ell + q - 1}{\ell + q - 1} \binom{\ell + q - 1}{\ell}, & \text{if } \ell \geq 1, \\ 1, & \text{if } \ell = 0. \end{cases}$$

We denote the class of all complex algebraic polynomials of degree at most n in $q + 1$ variables, restricted to \mathbb{S}^q , by Π_n^q . It is well known that

$$\Pi_n^q = \bigoplus_{\ell=0}^n \mathbf{H}_\ell^q,$$

where \bigoplus denotes the orthogonal direct sum. There exists a sequence of polynomials $\{\mathcal{P}_\ell(q+1; x)\}_{\ell=0}^\infty$, known as the Legendre polynomial of degree ℓ in \mathbb{R}^{q+1} , with the property that

$$\int_{-1}^1 \mathcal{P}_\ell(q+1; x) \mathcal{P}_k(q+1; x) w_q(x) dx = \frac{\omega_q}{\omega_{q-1} d_\ell^q} \begin{cases} 1, & \text{if } \ell = k, \\ 0, & \text{otherwise.} \end{cases}$$

The analogue of the reproduction formula (4.2.3) is the following *Funk-Hecke formula*. For any $\phi \in L^1_{w_q}[-1, 1]$, $\mathbf{y} \in \mathbb{S}^q$, and any $Y_\ell \in \mathbb{H}_\ell^q$, we have

$$\int_{\mathbb{S}^q} \phi(\mathbf{x} \cdot \mathbf{z}) Y_\ell(\mathbf{z}) d\mu_q(\mathbf{z}) = \omega_{q-1} Y_\ell(\mathbf{x}) \int_{-1}^1 \phi(t) \mathcal{P}_\ell(q+1; t) w_q(t) dt \quad (4.3.3)$$

$$=: \frac{\omega_q}{d_\ell^q} \hat{\phi}(\ell) Y_\ell(\mathbf{x}). \quad (4.3.4)$$

We remark that our definition of $\hat{\phi}$ in (4.3.4) is chosen so that the Legendre expansion of ϕ has the form $\sum \hat{\phi}(\ell) \mathcal{P}_\ell(q+1; \cdot)$.

Next, we make some observations regarding a set of scattered sites on the sphere. Let \mathcal{C} be a finite set of distinct points on \mathbb{S}^q ,

$$\delta_{\mathcal{C}} := \text{mesh}(\mathcal{C}) := \sup_{\mathbf{x} \in \mathbb{S}^q} \text{dist}(\mathbf{x}, \mathcal{C}),$$

where the distance is the geodesic distance on the sphere.

Definition 4.3.1 Let \mathcal{R} be a finite collection of closed, nonoverlapping (i.e., having no common interior points) regions $R \subset \mathbb{S}^q$ such that $\cup_{R \in \mathcal{R}} R = \mathbb{S}^q$. We will say that \mathcal{R} is \mathcal{C} -compatible if each $R \in \mathcal{R}$ is a spherical simplex containing at least one point of \mathcal{C} in its interior.

Unlike the case of $[-\pi, \pi]^s$, it is not immediately clear that one can choose a partition so that each region in this partition contains exactly one point of \mathcal{C} .

Theorem 4.3.1 There exists a \mathcal{C} -compatible decomposition \mathcal{R} for which each $R \in \mathcal{R}$ is a spherical simplex. By replacing \mathcal{C} by a subset, each $R \in \mathcal{R}$ contains a unique point $\xi \in \mathcal{C}$ in its interior. Further, for any $R \in \mathcal{R}$ and \mathbf{x} in the interior of R , any geodesic through \mathbf{x} intersects R in exactly two points.

We will assume in the remainder of this section that a \mathcal{C} -compatible decomposition \mathcal{R} is chosen, and the set \mathcal{C} is reduced so that there is a one-one correspondence $\xi \in \mathcal{C} \leftrightarrow R_\xi \in \mathcal{R}$, $\xi \in R_\xi$. Corresponding to this partition, we may define the following discrete norms for $f : \mathcal{C} \rightarrow \mathbb{C}$:

$$\|f\|_{\mathcal{C},p} := \begin{cases} \left(\sum_{\xi \in \mathcal{C}} |f(\xi)|^p \mu_q(R_\xi) \right)^{1/p}, & \text{if } 1 \leq p < \infty, \\ \sup_{\xi \in \mathcal{C}} \{|f(\xi)|\}, & \text{if } p = \infty. \end{cases} \quad (4.3.5)$$

The analogue of Theorem 3.3.2 and Theorem 3.3.1 is the following:

Theorem 4.3.2 There exist constants α_q and N_q with the following property. Let $1 \leq p \leq \infty$, \mathcal{C} be a finite set of distinct points on \mathbb{S}^q , and n be an integer with $N_q \leq n \leq \alpha_q \delta_{\mathcal{C}}^{-1}$. Then there exist nonnegative weights $\{a_\xi\}_{\xi \in \mathcal{C}}$, with

$$0 \leq a_\xi \leq c \mu_q(R_\xi), \quad (4.3.6)$$

such that for every $P \in \Pi_n^q$,

$$\frac{1}{\omega_q} \int_{\mathbb{S}^q} P(\mathbf{x}) d\mu_q(\mathbf{x}) = \sum_{\xi \in \mathcal{C}} a_\xi P(\xi), \quad (4.3.7)$$

and

$$\|P\|_{\mathcal{C},p} \sim \|P\|_{\mathbb{S}^q,p}. \quad (4.3.8)$$

Further,

$$|\{\xi : a_\xi \neq 0\}| \sim n^q \sim \dim(\Pi_n^q). \quad (4.3.9)$$

We remark that in contrast to Theorem 3.3.1, considerations involving one-sided approximations allow us to choose the weights a_ξ to be both nonnegative and satisfying (4.3.6). Let $\nu_{\mathcal{C}}$ be the measure that associates the mass a_ξ with each $\xi \in \mathcal{C}$. For continuous $f : \mathbb{S}^q \rightarrow \mathbb{R}$, we write

$$E_{\mathbb{S}^q,n,\infty}(f) := \min_{P \in \Pi_n^q} \|f - P\|_{\mathbb{S}^q,\infty}.$$

Let k_q be the smallest integer greater than $(q-1)/2$. The analogue of Theorem 3.4.1 is the following:

Theorem 4.3.3 *There exists a sequence of univariate polynomials τ_n^q of degree $(k_q + 2)n$ such that the operators defined by*

$$T_n^{\mathcal{C}}(f, \mathbf{x}) := \int_{\mathbb{S}^q} f(\mathbf{y}) \tau_n^q(\mathbf{x} \cdot \mathbf{y}) d\nu_{\mathcal{C}}(\mathbf{y}) \quad (4.3.10)$$

satisfy the following properties. For each $P \in \Pi_n^q$, $T_n^{\mathcal{C}}(P) = P$. For each $f : \mathcal{C} \rightarrow \mathbb{R}$, $T_n^{\mathcal{C}}(f) \in \Pi_{(k_q+2)n}^q$, and for $1 \leq p \leq \infty$,

$$\|T_n^{\mathcal{C}}(f)\|_{\mathbb{S}^q, p} \leq c \|f\|_{\mathbb{S}^q, p, \nu_{\mathcal{C}}}, \quad (4.3.11)$$

where the constant c is independent of \mathcal{C} and n . In particular, if $f \in C(\mathbb{S}^q)$ then

$$E_{\mathbb{S}^q, (k_q+2)n, \infty}(f) \leq \|f - T_n^{\mathcal{C}}(f)\|_{\mathbb{S}^q, \infty} \leq c E_{\mathbb{S}^q, n, \infty}(f). \quad (4.3.12)$$

Using the Funk-Hecke formula (4.3.3), we may now construct ZF networks based on the operators $T_n^{\mathcal{C}}$. Towards this end, let $\phi : [-1, 1] \rightarrow \mathbb{R}$ be a continuous function, which satisfies the additional condition

$$\hat{\phi}(\ell) \neq 0, \quad \ell = 0, 1, 2, \dots, \quad (4.3.13)$$

where $\hat{\phi}(\ell)$ is defined in (4.3.4). In view of the Funk-Hecke formula (4.3.3), the class of all ZF networks is not dense in $C(\mathbb{S}^q)$ without this condition; in fact, if $\hat{\phi}(\ell) = 0$ for some ℓ then this class is orthogonal to \mathbb{H}_{ℓ}^q . For integers $n = 0, 1, \dots$, we write

$$E_{n,p}(\phi) := \min \| \phi - P \|_{w_q, p}, \quad (4.3.14)$$

where the minimum is taken over all univariate polynomials P of degree not exceeding n .

For integer $N \geq 1$, we define the univariate polynomial ϕ_N^+ of degree N by

$$\phi_N^+ := \sum_{\ell=0}^N \frac{(d_{\ell}^q)^2}{\omega_q^2 \hat{\phi}(\ell)} \mathcal{P}_{\ell}(q+1; \cdot). \quad (4.3.15)$$

Next, we define the operators

$$\Phi f(\mathbf{x}) := \int_{\mathbb{S}^q} \phi(\mathbf{x} \cdot \mathbf{y}) f(\mathbf{y}) d\mu_q(\mathbf{y}), \quad f \in L^1(\mathbb{S}^q), \mathbf{x} \in \mathbb{S}^q, \quad (4.3.16)$$

$$\Phi_N^+ f(\mathbf{x}) := \int_{\mathbb{S}^q} \phi_N^+(\mathbf{x} \cdot \mathbf{y}) f(\mathbf{y}) d\mu_q(\mathbf{y}), \quad f \in L^1(\mathbb{S}^q), \mathbf{x} \in \mathbb{S}^q, \quad (4.3.17)$$

and for $f : \mathcal{C} \rightarrow \mathbb{R}$,

$$\Phi^{\mathcal{C}} f(\mathbf{x}) := \int_{\mathbb{S}^q} \phi(\mathbf{x} \cdot \mathbf{y}) f(\mathbf{y}) d\nu_{\mathcal{C}}(\mathbf{y}), \quad \mathbf{x} \in \mathbb{S}^q. \quad (4.3.18)$$

We observe that $\Phi^{\mathcal{C}}$ is a ZF network, obtained by a discretization of the operator Φ . Further, the relation (4.3.9) implies that the number of evaluations of ϕ involved in the computation of $\Phi^{\mathcal{C}} f$ is of the order of magnitude $\delta_{\mathcal{C}}^{-q}$.

The analogue of Theorem 4.2.1 is the following:

Theorem 4.3.4 *Let $1 \leq p \leq \infty$, \mathcal{C} be a set of distinct points on \mathbb{S}^q , M, N be integers that satisfy $N \geq N_q$ and $M + N \leq \alpha_q \delta_{\mathcal{C}}^{-1}$, $\phi \in C[-1, 1]$ satisfy (4.3.13). We write*

$$\beta := \beta(p) := \max\left(0, \frac{1}{p} - \frac{1}{2}\right), \quad m_N := \min_{0 \leq \ell \leq N} \frac{|\hat{\phi}(\ell)|}{d_{\ell}^q}. \quad (4.3.19)$$

Then for any $P \in \Pi_N^q$, we have

$$\|P - \Phi^{\mathcal{C}} \Phi_N^+ P\|_{\mathbb{S}^q, p} \leq c \frac{N^{\beta}}{m_N} E_{M, p}(\phi) \|P\|_{\mathbb{S}^q, p}. \quad (4.3.20)$$

We are now in a position to state our approximation theorems for the case of “arbitrary” functions, analogous to Theorem 4.2.2.

Theorem 4.3.5 *Let n, M be integers, $n \geq N_q$, and $1 \leq p \leq \infty$. Let \mathcal{C} be a set of distinct points on \mathbb{S}^q such that $(k_q + 2)n + M \leq \alpha_q \delta_{\mathcal{C}}^{-1}$. Let $N = (k_q + 1)n - 1$. If $f \in C(\mathbb{S})$, then*

$$\|f - \Phi^{\mathcal{C}} \Phi_N^+ T_n^{\mathcal{C}} f\|_{\mathbb{S}^q, \infty} \leq c \left(E_{\mathbb{S}^q, n, \infty}(f) + \frac{E_{M, \infty}(\phi)}{m_N} \|f\|_{\mathbb{S}^q, \infty} \right). \quad (4.3.21)$$

Theorem 4.3.4 leads in this case to the following “converse theorem”, relating the degree of approximation by polynomials to that by our operators.

Theorem 4.3.6 *We continue the notations and conditions of Theorem 4.3.5. In addition, let $R > 0$, $0 < \gamma \leq R$ and*

$$\frac{E_{M, \infty}(\phi) n^{\beta}}{m_N} \leq c n^{-R}. \quad (4.3.22)$$

If $f \in C(\mathbb{S}^q)$, and $\|f - \Phi^{\mathcal{C}} \Phi_N^+ T_n^{\mathcal{C}} f\|_{\mathbb{S}^q, \infty} \leq c n^{-\gamma}$ then $E_{\mathbb{S}^q, n, \infty}(f) \leq c n^{-\gamma}$.

5 Optimal Approximation by neural networks

5.1 Introduction

In our constructions of the last chapter, the emphasis was to obtain some estimates in the case of activation functions as general as possible. As the results showed, the smoother the activation function, the better the estimates. In this section, we show that neural networks with activation functions with more stringent conditions than the minimal necessary for density can be constructed to provide an optimal approximation order. Towards this end, we translate in the next section the results of the previous chapters to the context of algebraic polynomial approximation. In Section 5.3, we will explain the construction of neural networks from polynomials, and summarize by giving the results in the context of approximation of aperiodic functions. In Section 5.4, we explain how to do the constructions using scattered data.

5.2 Polynomial approximation

There is a natural connection between functions defined on $[-1, 1]$ and periodic functions. If $f : [-1, 1] \rightarrow \mathbb{R}$, then $f^\circ(\theta) = f(\cos \theta)$ defines an even periodic function. Conversely, if $f^\circ : \mathbb{R} \rightarrow \mathbb{R}$ is an even, 2π -periodic function, then the formula $f(\cos \theta) := f^\circ(\theta)$ defines a function on $[-1, 1]$. The continuity and differentiability properties are preserved under this correspondence. Let $C[a, b]$ denote the class of all continuous real valued functions on the interval $[a, b]$ with

$$\|f\|_{[a,b]} := \max_{x \in [a,b]} |f(x)|, \quad (5.2.1)$$

and, for integer $r \geq 0$, $W_r([a, b])$ denote the class of all r times continuously differentiable functions on $[a, b]$ with

$$\|f\|_{r,[a,b]} := \sum_{k=0}^r \|f^{(k)}\|_{[a,b]} \quad (5.2.2)$$

Then $f \in C[-1, 1]$ if and only if $f^\circ \in C^*$, and $\|f\|_{[-1,1]} = \|f^\circ\|^*$. Also, $f \in W_r([-1, 1])$ if and only if $f^\circ \in W_r^*$, and

$$\|f^{\circ(r)}\|^* \leq c \|f\|_{r,[-1,1]} \quad (5.2.3)$$

with an absolute constant c .

Next, if we write

$$T_n(\cos \theta) := \cos n\theta, \quad n = 0, 1, \dots, \quad (5.2.4)$$

then we have $T_0(x) = 1$, $T_1(x) = x$, and

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \dots.$$

Therefore, T_n is a (algebraic) polynomial of degree n , having leading coefficient 2^{n-1} for $n = 1, 2, \dots$. Let Π_n denote the class of all polynomials of degree at most n . Then $\{T_k\}_{k=0}^n$ is a basis for Π_n . Therefore, for any $P \in \Pi_n$, there exists an even trigonometric polynomial $P^\circ \in \mathbb{H}_n$ given by $P^\circ(\theta) := P(\cos \theta)$. Conversely, if $P^\circ \in \mathbb{H}_n$ is an even trigonometric polynomial, then there exists $P \in \Pi_n$ given by $P(\cos \theta) := P^\circ(\theta)$.

If

$$E_n(f) := \min_{P \in \Pi_n} \|f - P\|_{[-1,1]} \quad (5.2.5)$$

then it is easy to verify that $E_n(f) = E_n^*(f^\circ)$. Consequently, the Favard estimate (1.2.4) takes the form

$$E_n(f) \leq cn^{-r} \|f\|_{r,[-1,1]}. \quad (5.2.6)$$

Several improvements are possible in this result, but we will not go into these details. We observe that if $f \in C[-1, 1]$, then we may define

$$v_n(f, \cos \theta) := v_n^*(f^\circ, \theta). \quad (5.2.7)$$

Then $v_n(P) = P$ for every $P \in \Pi_n$, $v_n(f) \in \Pi_{2n-1}$, and

$$E_{2n-1}(f) \leq \|f - v_n(f)\|_{[-1,1]} \leq 4E_n(f).$$

In the multivariate case, we adopt the following notations. For a rectangle $R \subseteq \mathbb{R}^s$, we denote the class of all continuous real valued functions on R by $C(R)$, the class of all

functions having continuous partial derivatives in each variable up to order r by $W_r(R)$, and define

$$\|f\|_R := \max_{\mathbf{x} \in R} |f(\mathbf{x})|, \quad \|f\|_{r,R} := \sum_{0 \leq \mathbf{k} \leq r} \|D^{\mathbf{k}} f\|_R. \quad (5.2.8)$$

The correspondence $f \leftrightarrow f^\circ$ is given by

$$f(\cos \theta_1, \dots, \cos \theta_s) = f^\circ(\theta_1, \dots, \theta_s).$$

With this correspondence, $f \in C([-1, 1]^s)$ if and only if $f^\circ \in C_s^*$, $f \in W_r([-1, 1]^s)$ if and only if $f^\circ \in W_{r,s}^*$, and we have

$$\|f\|_{[-1,1]^s} = \|f^\circ\|_s^*, \quad \sum_{j=1}^s \|D_j^r f^\circ\|_s^* \leq c \|f\|_{r,[-1,1]^s}. \quad (5.2.9)$$

Let $\Pi_{n,s}$ denote the class of all algebraic polynomials in s variables having coordinate-wise degree at most n . We write

$$T_{\mathbf{n}}(\mathbf{x}) = \prod_{j=1}^s T_{n_j}(x_j), \quad \mathbf{x} = (x_1, \dots, x_s), \quad \mathbf{n} := (n_1, \dots, n_s). \quad (5.2.10)$$

The set $\{T_{\mathbf{k}}\}_{0 \leq \mathbf{k} \leq \mathbf{n}}$ is a basis for $\Pi_{n,s}$, and we have the correspondence $P \leftrightarrow P^\circ$, $P \in \Pi_{n,s}$, and even polynomials $P^\circ \in \mathbb{H}_{n,s}$ given by

$$P(\cos \theta_1, \dots, \cos \theta_s) = P^\circ(\theta_1, \dots, \theta_s).$$

Let

$$E_{n,s}(f) := \min_{P \in \Pi_{n,s}} \|f - P\|_{[-1,1]^s}. \quad (5.2.11)$$

Then $E_{n,s}(f) = E_{n,s}^*(f^\circ)$, and consequently, Theorem 1.4.1 gives

$$E_{n,s}(f) \leq cn^{-r} \|f\|_{r,[-1,1]^s}, \quad f \in W_r([-1, 1]^s). \quad (5.2.12)$$

We write (cf. (3.4.3))

$$v_{n,s}(f, (\cos \theta_1, \dots, \cos \theta_s)) := v_{n,s}^*(f^\circ, (\theta_1, \dots, \theta_s)). \quad (5.2.13)$$

Then (cf. Theorem 3.4.1) $v_{n,s}(P) = P$ for all $P \in \Pi_{n,s}$, $v_{n,s}(f) \in \Pi_{2n-1,s}$, and

$$E_{2n-1,s}(f) \leq \|f - v_{n,s}(f)\|_{[-1,1]^s} \leq 4E_{n,s}(f). \quad (5.2.14)$$

5.3 From polynomials to neural networks

Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be infinitely many times differentiable on an open interval $I \subseteq \mathbb{R}$. We assume that there exists a point $b \in I$ such that

$$\phi^{(k)}(b) \neq 0, \quad k = 0, 1, 2, \dots \quad (5.3.1)$$

Using the Baire category theorem, it is not too difficult to prove that this condition is always satisfied if ϕ is not a polynomial on any interval.

Next, we denote by $D_{\mathbf{w}}$ the operation of partial derivative with respect to \mathbf{w} . If \mathbf{w} is sufficiently close to $\mathbf{0}$ that $\mathbf{w} \cdot \mathbf{x} + b \in I$ for every $\mathbf{x} \in [-3, 3]^s$, then

$$D_{\mathbf{w}}^{\mathbf{k}} \phi(\mathbf{w} \cdot \mathbf{x} + b) = \mathbf{x}^{\mathbf{k}} \phi^{(|\mathbf{k}|)}(\mathbf{w} \cdot \mathbf{x} + b), \quad \mathbf{x} \in [-3, 3]^s,$$

where, in this section, $|\mathbf{k}| := \sum_{j=1}^s k_j$. (The reason for using $[-3, 3]^s$ in place of $[-1, 1]^s$ will be clearer in Section 5.4.) Hence,

$$\mathbf{x}^{\mathbf{k}} = \frac{1}{\phi^{(|\mathbf{k}|)}(b)} D_{\mathbf{w}}^{\mathbf{k}} \phi(\mathbf{w} \cdot \mathbf{x} + b) \Big|_{\mathbf{w}=\mathbf{0}}, \quad \mathbf{x} \in [-3, 3]^s. \quad (5.3.2)$$

We will approximate the partial derivative with a divided difference. For a function g of s variables and $h > 0$, we write

$$\begin{aligned} E_{h,j} &:= g(x_1, \dots, x_{j-1}, x_j + h, x_{j+1}, \dots, x_s), \\ \Delta_{h,j}^k g(\mathbf{x}) &:= (E_{h,j} - E_{-h,j})^k g(\mathbf{x}), \quad k = 1, 2, \dots, \\ \Delta_{h,j}^0 g(\mathbf{x}) &:= g(\mathbf{x}), \end{aligned}$$

and for $\mathbf{k} \in \mathbb{Z}^s$, $\mathbf{k} \geq 0$,

$$\Delta_h^{\mathbf{k}} g(\mathbf{x}) := \prod_{j=1}^s \Delta_{h,j}^{k_j} g(\mathbf{x}). \quad (5.3.3)$$

We write

$$\binom{\mathbf{k}}{\mathbf{r}} := \prod_{j=1}^s \binom{k_j}{r_j}, \quad \mathbf{k}! := \prod_{j=1}^s k_j!$$

It is elementary to verify that

$$\Delta_h^{\mathbf{k}} g(\mathbf{w}) = \sum_{0 \leq \mathbf{m} \leq \mathbf{k}} \binom{\mathbf{k}}{\mathbf{m}} (-1)^{\mathbf{k}-\mathbf{m}} g(\mathbf{w} + (2\mathbf{m} - \mathbf{k})h). \quad (5.3.4)$$

Further, using Taylor's theorem, it is not difficult to prove that if g has sufficiently many derivatives in a neighborhood of \mathbf{w} then for sufficiently small $h > 0$

$$\left| \frac{1}{(2h)^{|\mathbf{k}|} \mathbf{k}!} \Delta_h^{\mathbf{k}} g(\mathbf{w}) - D^{\mathbf{k}} g(\mathbf{w}) \right| \leq c_{\mathbf{k}}(g) h^2. \quad (5.3.5)$$

Thus, we have proved the following proposition (cf. Theorem 4.2.1).

Proposition 5.3.1 *Let ϕ be infinitely many times differentiable on an open interval $I \subseteq \mathbb{R}$, and there exist a point $b \in I$ such that (5.3.1) is satisfied. Let $\epsilon > 0$. For $h > 0$, we define*

$$\mathcal{N}_{\mathbf{k},h}(\mathbf{x}) := \frac{1}{(2h)^{|\mathbf{k}|} \mathbf{k}!} \sum_{0 \leq \mathbf{m} \leq \mathbf{k}} \binom{\mathbf{k}}{\mathbf{m}} (-1)^{\mathbf{k}-\mathbf{m}} \phi(h(2\mathbf{m} - \mathbf{k}) \cdot \mathbf{x} + b). \quad (5.3.6)$$

There exists $\delta(\phi, \mathbf{k}, \epsilon) > 0$ such that for $\mathbf{x} \in [-3, 3]^s$ and $0 < h < \delta(\phi, \mathbf{k}, \epsilon)$,

$$|\mathbf{x}^{\mathbf{k}} - \mathcal{N}_{\mathbf{k},h}(\mathbf{x})| \leq \epsilon. \quad (5.3.7)$$

In particular, if $P \in \Pi_{n,s}$, and $P(\mathbf{x}) = \sum_{0 \leq \mathbf{k} \leq n} a_{\mathbf{k}} \mathbf{x}^{\mathbf{k}}$, then by choosing a different value of ϵ (independent of P), we obtain that there exists $\delta(\phi, n, \epsilon) > 0$ such that

$$\|P - \sum_{0 \leq \mathbf{k} \leq n} a_{\mathbf{k}} \mathcal{N}_{\mathbf{k},h}\|_{[-3,3]^s} \leq \epsilon \|P\|_{[-1,1]^s}, \quad 0 < h < \delta(\phi, n, \epsilon). \quad (5.3.8)$$

The following theorem now follows easily.

Theorem 5.3.1 *Let ϕ be infinitely many times differentiable on an open interval $I \subseteq \mathbb{R}$, and there exist a point $b \in I$ such that (5.3.1) is satisfied. Let $f \in W_r([-1, 1]^s)$, and*

$$v_{n,s}(f, \mathbf{x}) =: \sum_{0 \leq \mathbf{k} \leq 2n-1} a_{\mathbf{k}}(f) \mathbf{x}^{\mathbf{k}}. \quad (5.3.9)$$

Then there exists $\delta := \delta(\phi, n)$ such that for all $h \in (0, \delta)$,

$$\|f - \sum_{0 \leq \mathbf{k} \leq 2n-1} a_{\mathbf{k}}(f) \mathcal{N}_{\mathbf{k},h}\|_{[-1,1]^s} \leq cn^{-r} \|f\|_{r,[-1,1]^s}. \quad (5.3.10)$$

We observe that the weights of the network $\sum_{0 \leq \mathbf{k} \leq 2n-1} a_{\mathbf{k}}(f) \mathcal{N}_{\mathbf{k},h}$ all belong to the set $\{(2\mathbf{m} - \mathbf{k})h : 0 \leq \mathbf{m}, \mathbf{k} \leq 2n-1\}$, having at most $(6n)^s$ elements. Therefore, the estimate (5.3.10) shows the construction of a neural network that provides optimal approximation to target functions in $W_r([-1, 1]^s)$. Moreover, the weights and thresholds of the network are selected independently of the target function. We can weaken the condition on ϕ to some extent. For example, one may allow ϕ to be in the uniform closure on a suitable cube of $\Pi_{\psi;k,1}$ for some ψ satisfying the conditions of the theorem and a fixed k .

5.4 Scattered data

It is possible to translate the scattered data operator $\tau_{n,s}^*$ to the aperiodic case, just as we translated the shifted average operator. For reasons that will be clearer later, we will do this translation in the case of data on $[-A, A]^s$ for $A > 0$. Analogously to the trigonometric case, we may define the mesh norm of a set $\mathcal{C} \subset [-A, A]^s$ by

$$\delta_{A,\mathcal{C}} := \max_{\mathbf{x} \in [-A,A]^s} \min_{\mathbf{y} \in \mathcal{C}} |\mathbf{x} - \mathbf{y}|_{\infty}. \quad (5.4.1)$$

For $\mathbf{x} = (x_1, \dots, x_s) \in [-A, A]^s$, we define $\theta_A(\mathbf{x}) = (\theta_1, \dots, \theta_s)$ by

$$x_j = A \cos \theta_j, \quad j = 1, \dots, s.$$

If \mathcal{C} is a set of distinct points in $[-A, A]^s$, we write

$$\mathcal{C}^\circ := \{\pm\theta_A(\mathbf{x}) : \mathbf{x} \in \mathcal{C}\}.$$

The translation of the relevant parts of Theorem 3.4.1 is the following.

Theorem 5.4.1 *Let $A > 0$, \mathcal{C} be a set of distinct points in $[-A, A]^s$ and $n \geq 1$ be an integer such that $\delta_{\mathcal{C}^\circ} < \pi/(2 \cdot 3^{s+4}n)$. In the following, all constants will depend upon A .*

(a) *There exist numbers $\{w_\xi\}_{\xi \in \mathcal{C}}$ such that*

$$|w_\xi| \leq cn^{-s}, \quad \xi \in \mathcal{C}, \quad (5.4.2)$$

and for every $P \in \Pi_{n,s}$,

$$\frac{1}{(2\pi)^s} \int_{[-A,A]^s} P(\mathbf{t}) \prod_{j=1}^s (A^2 - t_j^2)^{-1/2} dt = \sum_{\xi \in \mathcal{C}} w_\xi P(\xi). \quad (5.4.3)$$

(b) *Let*

$$V_{A,n,s}(\mathbf{x}, \mathbf{t}) := V_{n,s}^*(\theta_A(\mathbf{x}) - \theta_A(\mathbf{t})) + V_{n,s}^*(\theta_A(\mathbf{x}) + \theta_A(\mathbf{t})), \quad (5.4.4)$$

and the operator $\tau_{A,n,s}$ be defined for $f : \mathcal{C} \rightarrow \mathbb{C}$ by

$$\tau_{A,n,s}(f, \mathbf{x}) := \tau_{A,n,s}(\mathcal{C}; f, \mathbf{x}) := \sum_{\xi \in \mathcal{C}} w_\xi f(\xi) V_{A,n,s}(\mathbf{x}, \xi). \quad (5.4.5)$$

Then $\tau_{A,n,s}(P) = P$ for every $P \in \Pi_{n,s}$. Also, for $f \in C([-A, A]^s)$, $\tau_{A,n,s}(f) \in \Pi_{2n-1,s}$, and we have

$$\begin{aligned} \|\tau_{A,n,s}(f)\|_{[-A,A]^s} &\leq c\|f\|_{[-A,A]^s}, \\ E_{A,2n-1,s}(f) &\leq \|f - \tau_{A,n,s}(f)\|_{[-A,A]^s} \leq cE_{A,n,s}(f), \end{aligned} \quad (5.4.6)$$

where

$$E_{A,n,s}(f) := \min_{P \in \Pi_{n,s}} \|f - P\|_{[-A,A]^s}.$$

We observe that for $\theta \in [0, \pi]$,

$$\frac{2}{\pi^2} \theta^2 \leq 2 \sin^2(\theta/2) = 1 - \cos \theta \leq \theta^2/4.$$

Consequently, to ensure that $\delta_{\mathcal{C}^\circ} < \pi/(2 \cdot 3^{s+4}n)$, we need \mathcal{C} to have points concentrated more densely near the faces of the cube $[-A, A]^s$. However, the whole interest in having scattered data is that we have no control on where they may lie. To circumvent this problem we will assume in the remainder of this section that the samples of our target function are available on $[-1 - \eta, 1 + \eta]^s$ for some $\eta \in (0, 1/2)$, even though the approximation is

desired only on $[-1, 1]^s$. Theoretically, this can always be arranged. If $f \in W_r([-1, 1]^s)$, the Whitney extension theorem gives an extension of f to $W_r([-1 - 2\eta, 1 + 2\eta]^s)$ with

$$\|f\|_{r, [-1-2\eta, 1+2\eta]^s} \leq c(\eta) \|f\|_{r, [-1, 1]^s}.$$

We may work with this extension. We are not aware of any practical applications where the data is not available on a cube slightly larger than that where the approximation is desired.

If $x, y \in [-1 - \eta, 1 + \eta]$, and we find θ, φ such that $x = (1 + 2\eta) \cos \theta$ and $y = (1 + 2\eta) \cos \varphi$, then the mean value theorem implies that

$$(1 + 2\eta)|\theta - \varphi| \geq |x - y| \geq \sqrt{1 - (1 + \eta)^2 / (1 + 2\eta)^2} |\theta - \varphi| \geq \sqrt{\eta/2} |\theta - \varphi|. \quad (5.4.7)$$

Now, let \mathcal{C} be a set of distinct points on $[-1 - \eta, 1 + \eta]^s$, and $n \geq 1$ be an integer such that $\delta_{1+\eta, \mathcal{C}} \leq (\eta/2)^{1/2} \pi / (2 \cdot 3^{s+4} n)$. In view of (5.4.7), we may enlarge \mathcal{C} artificially to a set $\mathcal{C}_E \subset [-1 - 2\eta, 1 + 2\eta]^s$ by adding points from $[-1 - 2\eta, 1 + 2\eta]^s \setminus [-1 - \eta, 1 + \eta]^s$, so that $\delta_{\mathcal{C}_E} \leq \pi / (2 \cdot 3^{s+4} n)$.

Next, we find an infinitely many times differentiable function ψ which is equal to 1 on $[-1, 1]^s$ and equal to 0 on $\mathbb{R}^s \setminus [-1 - \eta, 1 + \eta]^s$. The quantity $\tau_{1+2\eta, n, s}(\mathcal{C}_E; \psi f, \mathbf{x})$ can then be evaluated only using the values of f on \mathcal{C} . We get the desired neural network approximation on $[-1, 1]^s$ by writing $\tau_{1+2\eta, n, s}(\mathcal{C}_E; \psi f, \mathbf{x})$ in terms on monomials and using Proposition 5.3.1 exactly as in the proof of Theorem 5.3.1.

6 Notes

The material in Chapter 1 is standard, and can be found either explicitly or essentially in the books of Lorentz [3] or Timan [9]. Our proof of Theorem 2.2.1 is taken from the treatise [10, Chapter X, Section 3] of Zygmund. The notion of nonlinear n -widths in Section 2.3 is introduced by Howard, DeVore and Micchelli in [1]. Our proof of Theorem 2.3.1 is an adaptation of the ideas in [1] and [3]. The material in Section 2.4 is standard, and can be found in [3] and [9]. The material in Chapter 3 is new, but is a simplified version of the arguments in [6] and [7]. Theorem 3.2.2 is in [2, p. 20]. The Whitney extension theorem is proved in [8, §VI.3.1]. Section 4.2 is based on [5], where the classical shifted average operators were used instead of the scattered data quasi-interpolation operators, which are only introduced here. The material in Section 4.3 is taken from [6] and [7]. The material in Chapter 5 is based on [4], although the thoughts on training with scattered data in Section 5.4 are new.

References

- [1] R. DEVORE, R. HOWARD AND C. A. MICCHELLI, *Optimal nonlinear approximation*, Manuscripta Mathematica, **63** (1989), 469-478.
- [2] R. B. HOLMES, “Geometric functional analysis and its applications”, Springer-Verlag, New York, 1975.
- [3] G. G. LORENTZ, “Approximation of Functions”, Holt, Rinehart and Winston, New York, 1966.
- [4] H. N. MHASKAR, *Neural networks for optimal approximation of smooth and analytic functions*, Neural Computation, **8** (1996), 164- 177.
- [5] H. N. MHASKAR AND C. A. MICCHELLI, *Degree of approximation by neural and translation networks with a single hidden layer*, Advances in Applied Mathematics, **16** (1995), 151-183.
- [6] H. N. MHASKAR, F. J. NARCOWICH, AND J. D. WARD, *Quadrature Formulas on Spheres Using Scattered Data*, To appear in Math. Comp. .
- [7] H. N. MHASKAR, F. J. NARCOWICH, AND J. D. WARD, *Approximation Properties of Zonal Function Networks Using Scattered Data on the Sphere*, Advances in Computational Mathematics, **11** (1999), 121–137.
- [8] E. M. STEIN, “Singular integrals and differentiability properties of functions”, Princeton Univ. Press, Princeton, 1970.
- [9] A. F. TIMAN, “Theory of Approximation of Functions of a Real Variable”, Macmillan Co., New York, 1963.
- [10] A. ZYGMUND, “Trigonometric Series”, Cambridge University Press, Cambridge, 1977.