September 7, 1999

# Approximation by polynomials

Christer O. Kiselman

Contents:



### Introductory remarks

These notes comprise the main part of a course on approximation theory presented at Uppsala University in the Fall of 1998, viz. the part on polynomial approximation. The material is mainly classical. As sources I used Cheney [1966], Dzjadyk [1977], Korovkin [1959], and Lorentz [1953], as well as papers listed in the bibliography. The emphasis is on explaining the main ideas behind the most important techniques.

The last part of the course was on rational approximation and is not included here. I followed mainly Cheney [1966, Chapter 5, pp. 150–167]. I also discussed Padé approximation briefly, following Cheney [1966, Chapter 5, pp. 173–177] and the introduction in Rudälv [1998].

I am grateful to Tsehaye Kahsu Araaya for remarks to the manuscript.

## 1. Introduction

What is approximation theory? Maybe at the end of this course we will be able to give a mature answer to this question. For the time being, let me just sketch a kind of formal framework for the problems we are going to study.

We have a metric space $X$ and we want to approximate a given element $x \in X$ by an element $a$ of some subset $A$ of $X$. The elements of $A$ are "nice" or "tractable" and we want to make the distance between $x$ and $a$ as small as possible: we call this "to make a good approximation of $x$ by elements of $A$."

A *metric space* $X$ is a set $X$ together with a metric $d$. That $d$ is a *metric* means that it is a function $d\colon X^2 \to \mathbf{R}$ which satisfies three properties: it is *positive definite:*

$$d(x,y) \geqslant 0 \text{ for all } x, y \in X \text{ with equality precisely when } x = y; \tag{1.1}$$

it is *symmetric:*

$$d(x,y) = d(y,x) \text{ for all } x, y \in X; \tag{1.2}$$

and it satisfies the *triangle inequality:*

$$d(x,z) \leqslant d(x,y) + d(y,z) \text{ for all } x,y,z \in X. \tag{1.3}$$

Any normed space gives rise to a metric space: we can define the distance between $x$ and $y$ as $d(x,y) = \|x-y\|$ or, more generally, as $d(x,y) = \|x-y\|^\alpha$ for any $\alpha$ satisfying $0 < \alpha \leqslant 1$.

To any element $x$ of $X$ and any subset $A$ of $X$ we associate the distance $d(x,A)$ from $x$ to $A$, which by definition is

$$d(x,A) = \inf_{a\in A} d(x,a), \qquad x \in X,\ A \subset X. \tag{1.4}$$

Obviously we have $0 \leqslant d(x,A) \leqslant +\infty$ with equality at the second place if and only if $A$ is empty and equality at the first place if and only if $a$ belongs to $\overline{A}$, the closure of $A$. So the elements $x$ such that $d(x,A) = 0$ are those which can be approximated arbitrarily well by nice elements. If $d(x,A) > 0$ there is a certain unavoidable error.

A very common situation is that we have an increasing sequence $(A_m)$ of sets whose union is dense in $X$, so that $d(x,A_m) \to 0$ as $m \to \infty$ for every $x \in X$. Then an interesting question is how fast the convergence is and how the rate of convergence depends on properties of the element $x$.

It may or may not happen that the infimum in (1.4) is a minimum. In other words, it may happen that there exists an element $a$, called *a best approximant*, such that

$$d(x,a) = d(x,A),$$

but it may also be the case that

$$d(x,a) > d(x,A) \text{ for all } a \in A.$$

In the latter case we are interested in constructing a sequence $(a_j)$ of elements of $A$ such that $d(x,a_j) \to d(x,A)$ as $j \to \infty$. We call such a sequence *an approximating sequence*. In the first case we may ask if there is a unique best approximant: the set

$$\{a \in A;\, d(x,a) = d(x,A)\},$$

may be empty, have exactly one element, or may have more than one element.

**Exercise 1.1.** Consider the spaces $l^p_n$, i.e., $\mathbf{R}^n$ with the norm

$$\|x\|_p = \left(\sum |x_j|^p\right)^{1/p} \text{ when } 1 \leqslant p < +\infty; \qquad \|x\|_\infty = \max |x_j|.$$

Let $A$ be a closed, convex and nonempty subset of $\mathbf{R}^n$. Prove that there is always a unique best approximant when $1 < p < +\infty$. Prove by examples that there does not necessarily exist a unique closest element in the remaining cases $p = 1, +\infty$. What about $0 < p < 1$?

A classical case when we always have a unique best approximant is the following.

**Theorem 1.2.** *Let $H$ be a real or complex Hilbert space and $A$ a nonempty, closed and convex subset of $H$. Then for any point $x \in H$ there is a unique closest point $\pi(x)$ in $A$. Any approximating sequence must tend to $\pi(x)$.*

*Proof.* We may take $H$ as a real Hilbert space. (Why?) After a translation we may assume that $x = 0$. Let $a_j$ be points in $A$ such that $d(0, a_j) = \|a_j\|$ tends to $d(0, A)$ when $j \to \infty$. We note the identity

$$\|\tfrac{1}{2}a_j + \tfrac{1}{2}a_k\|^2 + \|\tfrac{1}{2}a_j - \tfrac{1}{2}a_k\|^2 = \tfrac{1}{2}\|a_j\|^2 + \tfrac{1}{2}\|a_k\|^2,$$

which is easy to verify if we use the definition $\|x\| = \sqrt{(x\,|\,x)}$ of the norm in a Hilbert space in terms of the inner product $(x\,|\,y)$, $x, y \in H$. Then

$$\tfrac{1}{4}\|a_j - a_k\|^2 = \tfrac{1}{2}\|a_j\|^2 + \tfrac{1}{2}\|a_k\|^2 - \|\tfrac{1}{2}a_j + \tfrac{1}{2}a_k\|^2 \leqslant \tfrac{1}{2}\|a_j\|^2 + \tfrac{1}{2}\|a_k\|^2 - d(0, A)^2,$$

since $\frac{1}{2}a_j + \frac{1}{2}a_k$ belongs to $A$ in view of the convexity. We see that the right-hand side tends to zero as $j, k \to \infty$. This implies that $(a_j)$ is a Cauchy sequence, and it must therefore have a limit in $H$. The limit cannot depend on the sequence, for if we take two sequences $(a_j)$ and $(b_j)$ and mix them, the new sequence $(a_0, b_0, a_1, b_1, a_2, ...)$ must converge by the same argument. We call the limit $\pi(0)$; by translation we define $\pi(x) \in A$.

**Exercise 1.3.** Prove that the mapping $\pi\colon H \to A$ is continuous; more precisely that

$$\|\pi(x) - \pi(y)\| \leqslant \|x - y\|, \qquad x, y \in H.$$

Prove also that the set $A$ is contained in a half-space as soon as $x \notin A$: in the real case every $a \in A$ must satisfy $(a - x\,|\,\pi(x) - x) \geqslant \|\pi(x) - x\|^2$. What about the complex case?

**Exercise 1.4.** Prove that if $A$ is a closed linear subspace, then $x - \pi(x)$ is orthogonal to $\pi(x)$. Prove that we get two idempotent mappings $\pi$ and $I - \pi$, and determine all possible relations between the subspaces $\ker\pi$, $\ker(I - \pi)$, $\operatorname{im}\pi$, $\operatorname{im}(I - \pi)$.

So Hilbert space is an easy case where the best approximant is unique. However, there are other interesting cases when we can prove uniqueness of the best approximant.

Now we may perhaps dare to say that approximation theory is the study of approximating sequences, best approximants and their uniqueness or nonuniqueness in cases where $X$ is a space of interesting functions, and $A$ is some subspace of nice functions, like polynomials, trigonomentric polynomials,...

## 2. The Weierstrass approximation theorem

The best starting point for these lectures is the classical Weierstrass[1] approximation theorem. It says that for any continuous real-valued function $f$ on the interval $[0, 1]$ and any integer $k \geqslant 1$ there is a polynomial $p_k$ such that $|f(x) - p_k(x)| \leqslant 1/k$ for all

[1]Karl Theodor Wilhelm Weierstraß, 1815–1897, my advisor[5].

$x \in [0, 1]$. However, the degree of $p_k$ may have to be much larger than $k$; we shall return to that story.

Taylor polynomials can serve as approximants—if we are lucky. Given a function with derivatives of all orders at a point $a \in \mathbf{R}$, its *Taylor*[2] *polynomial* of degree $m$ at $a$ is

$$T_m(f)(x) = \sum_0^m f^{(j)}(a)(x-a)^j/j!, \qquad x \in \mathbf{R},\ m \in \mathbf{N}. \tag{2.1}$$

If the Taylor polynomials converge to $f$ on some interval $]a - r_1, a + r_2[$ containing $a$, then $f$ is the restriction to that interval of a holomorphic function defined in a disk in the complex plane, $\{z \in \mathbf{C};\ |z - a| < \max(r_1, r_2)\}$, thus a rather special function. In particular we see that if $f$ is identically zero near $a$, then we cannot hope for the Taylor polynomials to converge to $f$ at a point where $f$ is nonzero. To approximate continuous or even general $C^\infty$ functions we need to find other polynomials.

**Theorem 2.1** (The Weierstrass approximation theorem). *Let $f \in C([0, 1])$ and let $\varepsilon$ be a positive real number. Then there exists a polynomial $P$ such that $|f(x)-P(x)| < \varepsilon$ for all $x \in [0, 1]$.*

There are many generalizations of this theorem. For instance, we could let $f$ be complex-valued, or take values in a real or complex vector space of finite dimension. That is easy. We could let $f$ be a function of several real variables. Then it is also easy to formulate the result. And we could let $f$ be a function of several complex variables. That would require a more profound study, with skillful adaptions of both hypothesis and conclusion. Finally, of course, we may try to let the functions take their values in an infinite-dimensional space.

We start with a trivial identity:

$$1 = 1^m = (x + (1-x))^m = \sum_0^m \binom{m}{k} x^k (1-x)^{m-k}, \qquad x \in \mathbf{R},\ m \in \mathbf{N}. \tag{2.2}$$

Here, as usual,

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} \qquad m, k \in \mathbf{N},\ k \leqslant m, \tag{2.3}$$

is the binomial coefficient.[3] The formula holds for integers $m, k$ with $0 \leqslant k \leqslant m$. It is, however, often convenient to define the coefficient for all $k \in \mathbf{Z}$ and then set it equal to zero when $k$ is negative or larger than $m$. Indeed, there is no way you can choose $m + 1$ apples from $m$ apples. This convention allows us to work more easily with sums, whose limits need not always be indicated.

[2]Brook Taylor, 1685–1731.

[3]Read "$m$ choose $k$" in English, "$m$ super $k$" in Esperanto, and "$m$ över $k$" in Swedish.

What could we learn from (2.2) except that 1 can be written in many ways? Well, the function $x \mapsto \binom{m}{k} x^k(1-x)^{m-k}$ has a peak at $x = k/m$. The maximum is equal to

$$\binom{m}{k}\left(\frac{k}{m}\right)^k \left(1-\frac{k}{m}\right)^{m-k} = \binom{m}{k}\frac{k^k(m-k)^{m-k}}{m^m} \approx \sqrt{\frac{m}{2\pi k(m-k)}}.$$

The approximation holds when $k$ as well as $m-k$ are large, thus for $m$ large and $k/m$ near the middle of $[0,1]$. In particular, for $k \approx m/2$, the maximum is $\approx \sqrt{2/\pi m}$. We have used here Stirling's[4] formula, a simple version of which is

$$\Gamma(x) = (x-1)! \approx \sqrt{2\pi}e^{-x}x^{x-1/2}, \qquad x \gg 0. \tag{2.4}$$

More precisely, there is an asymptotic expansion valid as $|z|$ tends to $+\infty$ in any sector $|\arg z| \leqslant \pi - \varepsilon$, $\varepsilon > 0$, whose first terms are

$$\begin{aligned}\Gamma(z) &= (z-1)! \\ &= \sqrt{2\pi}e^{-z}z^{z-1/2}\left(1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{2488320z^4}+O(z^{-5})\right).\end{aligned} \tag{2.5}$$

For large $m$ the function $x^k(1-x)^{m-k}$ is pointed. However, it is not *very* pointed, in the following sense. Let us say that the *practical support* of a nonnegative function is the set where its value is at least $2^{-7}$ of its maximum, or perhaps the *precise practical support* where the value is at least $2^{-10}$ of the maximum. Then the practical support of $x^k(1-x)^{m-k}$ near $k = m/2$ is an interval of length approximately $3.12/\sqrt{m}$ (for the precise practical support, the interval is not much longer, viz. $3.72/\sqrt{m}$). The terms in the sum define a partition of unity over the interval $[0,1]$. However, it is not like the partitions we use in distribution theory when the supports are almost disjoint and one of the function is equal to 1 in a large compact set. In the sum (2.2) the supports overlap massively; using the notion of practical support again we see that $3.12\sqrt{m}$ of the terms overlap near $x = 1/2$. See the figures.

Anyway, it is reasonable to expect that

$$\sum f(k/m)\binom{m}{k}x^k(1-x)^{m-k}, \qquad x \in \mathbf{R},\ m \in \mathbf{N}^* = \mathbf{N} \smallsetminus \{0\},$$

which is a polynomial of degree $m$, would be a good approximation to $f$ on $[0,1]$. So let us define

$$B_m(f)(x) = \sum_0^m f(k/m)\binom{m}{k}x^k(1-x)^{m-k}, \qquad x \in \mathbf{R},\ m \in \mathbf{N}^*. \tag{2.6}$$

We shall call $B_m(f)$ *the Bernstein*[5] *polynomial* of $f$ of degree $m$.

[4] James Stirling, 1692–1770.

[5] Sergej Natanovič Bernstein, 1880–1968.

*Figure 1.* The polynomials $p_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ for $m = 10$, $k = 0, 1, ..., 10$.

We remark that the functions $x \mapsto \binom{m}{k} x^k (1-x)^k$ all have the same integral over $[0,1]$, viz. $1/(m+1)$. It follows that the integral of $B_m(f)$ is a Riemann sum for $f$:

$$\int_0^1 B_m(f)(x)dx = \frac{1}{m+1} \sum_0^m f(k/m). \tag{2.7}$$

**Proposition 2.2.** *The mapping $B_m\colon [0,1]^{\mathbf{R}} \to \mathcal{P}_m$ maps the real-valued functions on $[0,1]$ into the space $\mathcal{P}_m$ of polynomials of degree at most $m$, $m \in \mathbf{N}^*$. It is linear. It is increasing, i.e., $B_m(f) \leqslant B_m(g)$ if $f \leqslant g$; equivalently, it is positive, i.e., $B_m(f) \geqslant 0$ if $f \geqslant 0$. Moreover $B_m(1) = 1$ and $B_m(x)(x) = x$. Finally,*

$$B_m(x^2)(x) = x^2 + \frac{1}{m} x(1-x), \qquad x \in \mathbf{R},\ m \geqslant 1. \tag{2.8}$$

Thus the restriction of $B_m$ to the affine functions is the identity in $\mathcal{P}_1$. The restriction of $B_m$ to $\mathcal{P}_2$ is not exactly the identity, since if $p(x) = a_0 + a_1 x + a_2 x^2$ is a polynomial of degree at most 2, then

$$B_m(p)(x) = a_0 + \left(a_1 + \frac{a_2}{m}\right) x + \left(a_2 - \frac{a_2}{m}\right) x^2. \tag{2.9}$$

Thus $B_m$ maps $\mathcal{P}_2$ into itself and differs from the identity by an error which is nowhere larger than $|a_2|(4m)^{-1}$ and in particular tends to zero as $m \to +\infty$.

*Proof.* The linearity and positivity of $B_m$ is clear. The fact that $B_m(1) = 1$ is just the identity (2.2) which was the starting point of these deliberations. To see that

*Figure 2.* The polynomials $p_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ for $m = 100,\ k = 49, 50$. (Illustrating massive overlapping.)

$B_m(x)(x) = x$ we perform the following calculation.

$$\begin{aligned} B_m(x) &= \sum_1^m \binom{m}{k} \frac{k}{m} x^k (1-x)^{m-k} = \sum_1^m \binom{m-1}{k-1} x^k (1-x)^{m-k} \\ &= x \sum_0^{m-1} \binom{m-1}{j} x^j (1-x)^{m-1-j} = x B_{m-1}(1) = x, \qquad m \geqslant 2. \end{aligned}$$

The calculation supposes that $m \geqslant 2$, since $B_{m-1}$ is not defined otherwise; however, it is easy to verify that $B_1(x) = x$ separately.

Finally, for $f(x) = x^2$ we first perform the same trick as in the case of $B_m(x)$ above to get

$$B_m(x^2) = \sum_1^m \binom{m}{k} \frac{k^2}{m^2} x^k (1-x)^{m-k} = x \sum_0^{m-1} \frac{j+1}{m} \binom{m-1}{j} x^j (1-x)^{m-1-j}.$$

*Figure 3.* The polynomials $p_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ for $m = 1000$, $k = 499, 500$. (They are indistinguishable.)

Following Dzjadyk [1977:317] we now divide the last expression into two: the first with $j$, the second with 1, thus $B_m(x^2)$ is equal to

$$x\frac{m-1}{m}\sum_1^{m-1}\frac{j}{m-1}\binom{m-1}{j}x^j(1-x)^{m-1-j} + \frac{x}{m}\sum_0^{m-1}\binom{m-1}{j}x^j(1-x)^{m-1-j}$$
$$= x\Big(1-\frac{1}{m}\Big)B_{m-1}(x)(x) + \frac{x}{m}B_{m-1}(1)(x) = x^2 + \frac{1}{m}x(1-x), \qquad m \geqslant 2.$$

Here we need $m \geqslant 2$ to give a sense to $B_{m-1}$. However, $B_m(x^2)(x) = x = x^2 + \frac{1}{m}x(1-x)$ for $m = 1$, so (2.8) holds also then. This proves the proposition.

*Proof of Theorem 2.1.* This proof is due to Bernstein [1912]. Let $f$ be any continuous function on $[0, 1]$ and let $p$ be a polynomial of degree 2 which majorizes $f$. Then

$B_m(f) \leqslant B_m(p) = p + a_2 x(1-x)/m \leqslant p + |a_2|/(4m)$, where $a_2$ is the coefficient for $x^2$ in $p$. So $p + |a_2|/(4m)$ is a majorant of $B_m(f)$. Now a continuous function is the infimum of all its majorants of degree 2. More precisely, if a number $\varepsilon > 0$ is given, we can find a $\delta > 0$ such that $f(x) \leqslant f(x_0) + \varepsilon$ when $x$ and $x_0$ are any two points with $|x - x_0| \leqslant \delta$, for $f$ is uniformly continuous. Also $f$ is bounded on $[0,1]$, say $|f| \leqslant M$. Then $p(x) = f(x_0) + \varepsilon + a_2|x - x_0|^2$ is a majorant of $f$ if we choose $a_2 = 2M/\delta^2$. Taking the Bernstein polynomials of $f$ and $p$ we get

$$B_m(f) \leqslant B_m(p) = p + \frac{a_2}{m} x(1-x) \leqslant p + \frac{a_2}{4m}.$$

At the point $x_0$ the values are

$$B_m(f)(x_0) \leqslant p(x_0) + \frac{a_2}{4m} = f(x_0) + \varepsilon + \frac{a_2}{4m}.$$

For large $m$ we get

$$B_m(f)(x_0) \leqslant f(x_0) + 2\varepsilon. \tag{2.10}$$

In fact, we have to choose $m \geqslant \frac{1}{4}a_2\varepsilon^{-1} = \frac{1}{2}M\delta^{-2}\varepsilon^{-1}$. Therefore (2.10) holds for a large $m$ and for all $x_0$. By the same argument we get an estimate from below: $B_m(f) \geqslant f - 2\varepsilon$. Thus $|B_m(f) - f| \leqslant 2\varepsilon$ and the theorem is proved.

While we are at it, we can easily generalize the proof to other operators than $B_m$.

**Theorem 2.3.** *Let $A_m: C(I) \to C(I)$, $m \in \mathbf{N}^*$, where $I = [0,1]$, be positive linear operators such that $A_m(1) = 1$, $A_m(x)(x) = x + \alpha_m(x)$, and $A_m(x^2)(x) = x^2 + \beta_m(x)$. Then*

(2.11)

$$\|A_m(f) - f\|_\infty \leqslant \omega(\delta)\left(1 + \delta^{-2}\sup_{y \in I}|\beta_m(y) - 2y\alpha_m(y)|\right), \qquad \delta > 0,\ m \in \mathbf{N}^*,$$

*where $\omega$ is the modulus of continuity of $f$:*

$$\omega(\delta) = \sup_{x,y \in I}\big(|f(x) - f(y)|;\ |x - y| \leqslant \delta\big).$$

*Proof.* When $|x - y| \leqslant \delta$ we have $f(x) - f(y) \leqslant \omega(\delta)$ by definition. When $|x - y| > \delta$ we can divide the interval between $x$ and $y$ into not more than $1 + |x - y|/\delta$ intervals of length at most $\delta$, so that

$$f(x) - f(y) \leqslant \omega(\delta)(1 + |x - y|/\delta).$$

Thus in all cases

$$f(x) - f(y) \leqslant \omega(\delta)(1 + \delta^{-2}(x - y)^2), \qquad x, y \in I,\ \delta > 0.$$

Here the right-hand side is a polynomial of degree at most two in $x$ for fixed $y$, so if we apply the mapping $A_m$ in the variable $x$ we obtain

$$A_m(f)(x) - f(y) \leqslant \omega(\delta)\big[1 + \delta^{-2}\big(x^2 + \beta_m(x) - 2y(x + \alpha_m(x)) + y^2\big)\big].$$

We now choose $x = y$ and get

$$\begin{aligned} A_m(f)(y) - f(y) &\leqslant \omega(\delta)\big(1 + \delta^{-2}(\beta_m(y) - 2y\alpha_m(y))\big) \\ &\leqslant \omega(\delta)\big(1 + \delta^{-2}\sup_{y \in I}|\beta_m(y) - 2y\alpha_m(y)|\big). \end{aligned}$$

We now apply this result to $-f$ to get an inequality in the other direction. Combining the two we get (2.11) as claimed.

**Corollary 2.4.** *For any continuous function $f$ with modulus of continuity $\omega$, the Bernstein polynomials $B_m(f)$ satisfy*

$$\|B_m(f)-f\|_\infty \leqslant \tfrac{5}{4}\omega(1/\sqrt{m}), \qquad m\in\mathbf{N}^*.$$

*Proof.* With the notation of the theorem we have in this case $\alpha_m=0$, $|\beta_m(x)|\leqslant (4m)^{-1}$, so that the right-hand side of (2.11) becomes $\omega(\delta)(1+(4m\delta^2)^{-1})$. We just have to choose $\delta=1/\sqrt{m}$ to complete the proof.

**Exercise 2.5.** We have seen that $\|B_m(f)-f\|_\infty \leqslant C\omega(1/\sqrt{m})$, where the best constant $C$ is at most $5/4$. Prove that the best constant cannot be smaller than one. Thus $1\leqslant C\leqslant 5/4$; it appears that the exact value of $C$ is not known.

The Bernstein polynomials do not give the best possible approximation: we shall see that there are polynomials $P$ of degree $m$ such that $\|P-f\|_\infty\leqslant C\omega(1/m)$.

**Exercise 2.6.** Prove that if $A_m$, $m\in\mathbf{N}^*$, are positive linear operators such that $A_m(1)(x)=1+\alpha_m(x)$, $A_m(x)(x)=x+\beta_m(x)$, and $A_m(x^2)(x)=x^2+\gamma_m(x)$, where $\alpha_n$, $\beta_m$, and $\gamma_m$ tend to zero uniformly as $m\to\infty$, then $A_m(f)\to f$ uniformly for all $f\in C(I)$.

**Exercise 2.7.** Prove that $B_m(f)(x)\to f(x)$ if $f$ is bounded on $[0,1]$ and continuous only at the point $x$.

**Exercise 2.8.** Prove the formula $B_m(f)'=B_{m-1}(g)$, where

$$g(x)=m\left[f(\tfrac{m-1}{m}x+\tfrac{1}{m})-f(\tfrac{m-1}{m}x)\right],$$

a difference quotient of $f$ adjusted so that only values of the argument in $[0,1]$ appear. Deduce that if $f$ is a polynomial of degree at most $q$, then $B_m(f)$ is of degree at most $q$ (and thus of degree $\leqslant\min(m,q)$). Generalize to higher order difference quotients.

**Exercise 2.9.** Now do everything in several variables: define for $f\in C([0,1]^n)$,

$$B_m(f)(x)=\sum_{k\in\mathbf{N}^n} f(k/m)\binom{m}{k}x^k(1-x)^{m-k}, \qquad x\in\mathbf{R}^n,\ m\in(\mathbf{N}^*)^n,$$

where $\binom{m}{k}$, $k/m$ and $x^k$ have to be suitably defined. Prove that these Bernstein polynomials converge uniformly to $f\in C([0,1]^n)$.

## 3. Estimates for the Bernstein polynomials

We shall now take a closer look at the Bernstein polynomials, essentially following Lorentz [1953]. Let us define

$$p_{mk}(x)=\binom{m}{k}x^k(1-x)^{m-k}, \qquad m\in\mathbf{N}^*,\ 0\leqslant k\leqslant m, \tag{3.1}$$

and

$$T_{ms}(x)=\sum_{k=0}^{m}(k-mx)^s p_{mk}(x), \qquad m\in\mathbf{N}^*,\ s\in\mathbf{N}. \tag{3.2}$$

Since $p_{mk}$ has degree $m$, it follows that $T_{ms}$ has degree at most $m+s$. However, we shall see that $T_{ms}$ is of degree at most $s$ in $x$: a lot of cancellation takes place. Even more remarkable is the fact that $T_{ms}(x)$ is a polynomial also in $m$. We shall determine its degree.

To understand the role of the $T_{ms}$ let us remark that $B_m(f) = \sum f(k/m)p_{mk}$ and that, if we define $u_{as}(x) = (x-a)^s$, $x \in [0,1]$, $s \in \mathbf{N}$, then

$$B_m(u_{as})(x) = \sum u_{as}(k/m)p_{mk}(x) = m^{-s}\sum (k-ma)^s p_{mk}(x),$$

so that

$$B_m(u_{as})(a) = m^{-s}T_{ms}(a), \qquad a \in [0,1],\ s \in \mathbf{N}. \tag{3.3}$$

We can expect $B_m(u_{as})(a)$ to be smaller than $B_m(u_{as})(x)$ when $x$ is at some distance from $a$, since $|u_{as}|$ has its minimum at $a$.

We have actually already studied $T_{m0}, T_{m1}$ and $T_{m2}$. Indeed,

$$T_{m0}(x) = 1, \quad T_{m1} = 0, \quad \text{and } T_{m2}(x) = mx(1-x).$$

The first formula is just (2.2); the other two follow from Proposition 2.2. For larger values of $s$ we can state the following.

**Proposition 3.1.** *For a fixed $s \in \mathbf{N}$, $T_{ms}$ is a polynomial in $x$ and $m$. Its degree in $x$ is at most $s$. Its degree in $m$ is $\lfloor \frac{1}{2}s \rfloor$ (the integer part of $\frac{1}{2}s$), thus equal to $\frac{1}{2}s$ when $s$ is even and equal to $\frac{1}{2}(s-1)$ when $s$ is odd.*

*Proof.* We have

$$\begin{aligned} T'_{ms}(x) &= -msT_{m,s-1}(x) + \sum_0^m (k-mx)^s p'_{mk}(x) \\ &= -msT_{m,s-1}(x) + \sum_0^m (k-mx)^{s+1}\binom{m}{k}x^{k-1}(1-x)^{m-k-1} \\ &= -msT_{m,s-1}(x) + \frac{1}{x(1-x)}T_{m,s+1}. \end{aligned}$$

We rewrite this as a recursion formula

$$T_{m,s+1}(x) = x(1-x)[T'_{ms}(x) + msT_{m,s-1}(x)],$$

from which we can read off that the degrees in $x$ and $m$ are at most what is stated, starting from what we already know about the polynomials for $s = 0, 1, 2$. It remains to be seen that the degree in $m$ is not less than $\lfloor \frac{1}{2}s \rfloor$. However, it is possible to keep track of the coefficient of $m^s$ in $T_{m,2s}$ and $T_{m,2s+1}$ (the leading coefficient) and show that it does not vanish.

**Exercise 3.2.** Show that

$$\sum_{s=0}^{\infty} \frac{1}{s!}T_{ms}(x)y^s = [e^{-xy}(1-x+xe^y)]^m.$$

From this formula various properties of the $T_{ms}$ can be deduced.

Since the degree in $m$ of $T_{m,2s}$ is $s$, there exists, for each $s \in \mathbf{N}$, a constant $A_s$ such that

$$0 \leqslant T_{m,2s}(x) \leqslant A_s m^s, \qquad x \in [0,1],\ m \in \mathbf{N}^*.$$

This also yields an estimate for $B_m(u_{a,2s})(a)$:

$$0 \leqslant B_m(u_{a,2s})(a) = m^{-2s} T_{m,2s}(a) \leqslant A_s m^{-s}. \tag{3.4}$$

**Theorem 3.3.** *Let $f$ be a bounded function on $[0,1]$ such that $f'$ exists near a given point $a$ in that interval and such that $f''(a)$ exists. Then*

$$B_m(f)(a) = f(a) + \frac{a(1-a)}{2m} f''(a) + o(1/m), \qquad m \to \infty. \tag{3.5}$$

*Proof.* We already know that the result is true when $f$ is a polynomial of degree at most 2. For every positive $\varepsilon$ there is a constant $M$ such that

$$f(x) \leqslant p(x) + \varepsilon(x-a)^2 + M(x-a)^4, \qquad x \in [0,1],$$

where $p$ is the polynomial of degree at most 2 which agrees with the derivatives up to order 2 of $f$ at the point $a$. Then by (2.9),

$$B_m(f)(x) \leqslant p(x) + \varepsilon(x-a)^2 + \frac{x(1-x)}{2m}(p''(a) + 2\varepsilon) + MB_m(u_{a4})(x), \qquad x \in [0,1].$$

In particular, taking $x = a$, we obtain

$$B_m(f)(a) \leqslant p(a) + \frac{a(1-a)}{2m}(p''(a) + 2\varepsilon) + MB_m(u_{a4})(a), \qquad a \in [0,1].$$

Thus, in view of (3.4),

$$m(B_m(f)(a) - f(a)) \leqslant \frac{a(1-a)}{2}(f''(a) + 2\varepsilon) + MA_2 m^{-1}, \qquad a \in [0,1],$$

and we can conclude that

$$\varlimsup_{m\to\infty} m(B_m(f)(a) - f(a)) \leqslant \frac{a(1-a)}{2}(f''(a) + 2\varepsilon).$$

Since $\varepsilon$ is arbitrarily small, we get an inequality which, together with the inequality in the opposite direction, proves (3.5).

Thus for smooth functions the error $\|B_m(f) - f\|_\infty$ behaves like $C/m$ (and no better, unless of course $f \in \mathcal{P}_1$), whereas it is like $\omega(1/\sqrt{m}) \geqslant C/\sqrt{m}$ (and no better) for general continuous functions.

**Exercise 3.4.** Let $f$ be a smooth function such that lots of derivatives vanish at a particular point $a$: $f''(a) = f'''(a) = \cdots = f^{(q)}(a) = 0$. Prove that then $B_m(f)(a)$ converges to $f(a)$ fast.

Let us also consider the function $\psi_{a\delta}$ defined by $\psi_{a\delta}(x) = 0$ when $x \in [0,1]$, $|x-a| < \delta$, and $\psi_{a\delta}(x) = 1$ elsewhere in $[0,1]$. We can compare it with $u_{a,2s}$:

$$0 \leqslant \psi_{a\delta} \leqslant \delta^{-2s} u_{a,2s}, \qquad a \in [0,1],\ \delta > 0,\ s \in \mathbf{N}.$$

This gives

$$B_m(\psi_{a\delta})(x) \leqslant \delta^{-2s} B_m(u_{a,2s})(x), \qquad x, a \in [0,1],\ s \in \mathbf{N},$$

and, in particular at the point $a$,

$$0 \leqslant B_m(\psi_{a\delta})(a) \leqslant \delta^{-2s} B_m(u_{a,2s})(a) = m^{-2s}\delta^{-2s} T_{ms}(a), \qquad a \in [0,1],\ s \in \mathbf{N}.$$

This means that we have proved a useful estimate

$$\begin{aligned} \sum_{|k/m-a| \geqslant \delta} p_{mk}(a) &= \sum_{k=0}^{m} \psi_{a\delta}(k/m) p_{mk}(a) = B_m(\psi_{a\delta})(a) \\ &\leqslant \delta^{-2s} B_m(u_{a,2s})(a) = m^{-2s}\delta^{-2s} T_{m,2s}(a). \end{aligned} \tag{3.6}$$

Thus estimates for $T_{m,2s}$ will yield estimates for sums of $p_{mk}$ over those $k$ that satisfy $|k/m - a| \geqslant \delta$. For example it follows from (3.4) that, for any given positive $\delta$ and any $s \in \mathbf{N}$,

$$B_m(\psi_{a\delta})(a) \leqslant m^{-2s}\delta^{-2s} T_{m,2s}(a) \leqslant A_s \delta^{-2s} m^{-s}. \tag{3.7}$$

**Exercise 3.5.** Calculate $B_m(f)(\frac{1}{2})$ for $f(x) = |x - \frac{1}{2}|$, $x \in [0,1]$. (The answer should be compared with Bernstein's result [1913] that for the best polynomial approximation of $f$ the error is $O(1/m)$ and no better; as well as with Stahl's result [1992] on rational approximation that $E_{mm}(|x|; [-1,1]) = 8e^{-\pi\sqrt{m}}(1 + o(1))$, $m \to \infty$. Here $E_{mm}$ denotes the error when approximating by rational functions whose numerator and denominator are of degree at most $m$.)

## 4. Weierstrass' original proof

We shall take a look at Weierstrass' original proof of his theorem, Weierstrass [1885]. He used convolution by a kernel which is an entire function, and then the Taylor polynomial of the convolution product. So to describe this we have to define convolution.

Let $f$ and $g$ be two functions defined on all of $\mathbf{R}^n$. Then their *convolution product* is denoted by $f * g$ and defined by

$$(f * g)(x) = \int_{\mathbf{R}^n} f(y) g(x-y) dy, \qquad x \in \mathbf{R}^n, \tag{4.1}$$

whenever the integral has a sense. We can for instance take $f \in C_0(\mathbf{R}^n), g \in C(\mathbf{R}^n)$ or the other way around, i.e., we take two continuous functions (so that the integral over a bounded set exists) and then require that one of them have compact support (so that the integral over all of $\mathbf{R}^n$ exists). Or we can impose some growth at infinity. More advanced integrals can be used; it is proved in Lebesgue integration theory that (4.1) exists for almost all $x \in \mathbf{R}^n$ if $f, g \in L^1(\mathbf{R}^n)$, and that the convolution product is then itself an element of $L^1(\mathbf{R}^n)$. But in that case the integral in (4.1) may not exist for certain values of $x$.

Given now $f \in C(I)$, $I = [0,1]$, we first have to extend it to the whole axis. This is easy: we just put $f(x) = f(0)$ for negative $x$ and $f(x) = f(1)$ for $x > 1$. The new function is then an element of $C(\mathbf{R})$. Or we can extend it to an element of $C_0(\mathbf{R})$ by declaring it to be zero for $x \leqslant -\eta$ as well as for $x \geqslant 1+\eta$, and then define it as a first-degree polynomial in $[-\eta, 0]$ and $[1, 1+\eta]$; thus $f \in C_0(\mathbf{R})$. Here $\eta$ is a fixed positive number. Either way the integral

$$(W_m * f)(x) = \int_{\mathbf{R}} W_m(y) f(x-y)\,dy, \qquad x \in \mathbf{R},$$

exists if we let $W$ be the function

$$W(x) = Ce^{-x^2}, \qquad x \in \mathbf{R}, \tag{4.2}$$

with the constant $C$ determined such that

$$\int_{\mathbf{R}} W(x)\,dx = 1,$$

and

$$W_m(x) = mW(mx), \qquad x \in \mathbf{R},\ m \in \mathbf{N}^*.$$

It is now a well-known result that $W_m * f$ converges to $f$ uniformly. Indeed, we have the following theorem.

**Theorem 4.1.** *Let $K_j$, $j \in \mathbf{N}^*$, be integrable functions on the real axis which satisfy:*
(a) $\sup_j \int_{\mathbf{R}} |K_j|\,dx < +\infty$*;*
(b) $\int_{\mathbf{R}} K_j\,dx \to 1$ *as* $j \to \infty$*;*
(c) $\int_{|x|>\delta} |K_j|\,dx \to 0$ *as* $j \to \infty$ *for every positive number* $\delta$*.*

*Then* $K_j * f \to f$ *uniformly as* $j \to \infty$ *for every* $f \in C_0(\mathbf{R})$*.*

If $K_j \geqslant 0$ we can simplify the hypotheses to: $\int_{\mathbf{R}} K_j\,dx \to 1$ and $\int_{|x|>\delta} K_j\,dx \to 0$.

**Exercise 4.2.** Prove this theorem. Determine the constant $C$ in (4.2).

Obviously $K_m = W_m$ satisfies the hypothesis of the theorem, so we get a sequence of continuous functions $f_m = W_m * f$ which converges to $f \in C_0(\mathbf{R})$.

However, $W_m$ is not only a continuous function on the real axis; it is also the restriction to $\mathbf{R}$ of an entire function, i.e., a function which is holomorphic in the whole complex plane.

**Lemma 4.3.** *Let $W(z) = Ce^{-z^2}$, $z \in \mathbf{C}$, and define*

$$(W * f)(z) = \int_{\mathbf{R}} W(z-y)f(y)dy = \int_{-\eta}^{1+\eta} W(z-y)f(y)dy, \qquad z \in \mathbf{C},$$

*for any function $f \in C_0(\mathbf{R})$ with support in $[-\eta, 1+\eta]$ for a fixed positive $\eta$. Then $W * f$ is holomorphic everywhere in $\mathbf{C}$.*

*Sketch of proof.* We have

$$W(z-y) = \sum_{j \in \mathbf{N}} c_j (z-y)^j$$

with uniform convergence when $|z-y| \leqslant R$ and $R$ is any positive number. Therefore we can integrate $W(z-y)f(y)$ with respect to $y$ termwise:

$$(W * f)(z) = \sum_{j \in \mathbf{N}} c_j \int_{-\eta}^{1+\eta} (z-y)^j f(y)dy.$$

Now $\int_{-\eta}^{1+\eta}(z-y)^j f(y)dy$ is a polynomial in $z$, thus an entire function. The series converges uniformly for $|z| \leqslant R - 1 - \eta$ and its sum must therefore be holomorphic for $|z| < R - 1 - \eta$, hence everywhere.

**Exercise 4.4.** Complete the proof with all details.

**Exercise 4.5.** Give an alternative proof that $W*f$ is entire by studying the difference quotient

$$\frac{(W * F)(a+z) - (W * F)(a)}{z}$$

and proving that it converges to $(W' * f)(a)$ as $z \to 0$, where $W'$ denotes the complex derivative of $W$, $W'(z) = (d/dz)Ce^{-z^2} = -2Cze^{-z^2}$.

Now for entire functions the Taylor polynomials (see (2.1)) serve very well as approximants. We thus take the Taylor polynomials of $W_m * f$ to a degree which is so large that the error on the interval $[-\eta, 1+\eta]$ is not more than $1/m$. This proves the theorem.

Landau[6] [1908] gave a similar proof: he took $K_j$ as a polynomial $L_j$ of degree $2j$:

$$L_j(x) = \alpha_j \left(1 - (x/c)^2\right)^j, \qquad x \in \mathbf{R},$$

where $c$ is a fixed number larger than 1, and the constant $\alpha_j$ is chosen so that

$$\alpha_j \int_{-c}^{c} \left(1 - (x/c)^2\right)^j dx = 1.$$

We assume that $f \in C_0(\mathbf{R})$ has its support in $[-\eta, 1+\eta]$, where $1 + 2\eta = c$. We see that $K_j * f = L_j * f$ is a polynomial in this case. Landau's proof is thus like

[6]Edmund Georg Hermann Landau, 1877–1938.

Weierstrass' except that the last step, when we approximated an entire function by polynomials, is unnecessary.

## 5. The Stone–Weierstrass approximation theorem

An *order relation* in a set $X$ is a relation (a subset of $X^2$) which satisfies three conditions: it is *reflexive*, *antisymmetric* and *transitive*. This means, if we denote the relation by $\leqslant$, that for all $x, y, z \in X$,

$$x \leqslant x;$$

$$x \leqslant y \text{ and } y \leqslant x \text{ implies } x = y;$$

$$x \leqslant y \text{ and } y \leqslant z \text{ implies } x \leqslant z.$$

An *ordered set* is a set $X$ together with an order relation.

A *lattice* is an ordered set $L$ where each set consisting of two elements has an infimum and a supremum (a greatest minorant and a least majorant). This means that for any two $x, y \in L$ there exists $a$ and $b$ in $L$ such that $a \leqslant x, y$, and such that $z \leqslant x, y$ implies $z \leqslant a$; and similarly for $b$: $x, y \leqslant b$, and $x, y \leqslant z$ implies $b \leqslant z$. The supremum and infimum are of course unique. It is customary to denote the infimum and supremum of two elements $x, y$ by $x \wedge y$ and $x \vee y$, respectively.

A *sublattice* is defined just like a subgroup: that $L_1$ is a sublattice of $L$ means that for all $x, y \in L_1$, $x \wedge y$ and $x \vee y$, when calculated in $L$, are elements of $L_1$.

A lattice is said to be *complete* if any subset has an infimum and a supremum.

*Example 5.1.* The space of real-valued continuous functions on a topological space is a lattice. The space $C^1(\mathbf{R}^n)$ of continuously differentiable functions on $\mathbf{R}^n$ is not a sublattice of $C(\mathbf{R}^n)$. It is not even a lattice on its own.

*Example 5.2.* The set $[-\infty, +\infty]^{\mathbf{R}^n}$ of all functions defined on $\mathbf{R}^n$ and with values in the extended real line $[-\infty, +\infty] = \mathbf{R} \cup \{-\infty, +\infty\}$ is a lattice under the usual order for real numbers, extended in an obvious way to the two infinities. The subset of all convex functions is ordered in the same way, and is also a lattice under this order. However, the convex functions $CVX(\mathbf{R}^n)$ do not form a sublattice of $[-\infty, +\infty]^{\mathbf{R}^n}$. The supremum of two convex functions is equal to the pointwise supremum of them:

$$f \vee g = \max(f, g),$$

but the infima are different in the two lattices: the infimum in the lattice of convex functions is

$$f \wedge_{cvx} g = \sup \big[h; h \in CVX(\mathbf{R}^n), h \leqslant f, g\big],$$

where the supremum is calculated in $[-\infty, +\infty]^{\mathbf{R}^n}$ and has a sense because that lattice is complete.

**Lemma 5.3.** *Let $A$ be a closed subalgebra of $C(K)$, the algebra of all real-valued continuous functions on a compact topological space $K$. Then $A$ is a sublattice of $C(K)$.*

*Proof.* In $C(K)$ the lattice operations can be defined in terms of the absolute value:

$$f \wedge g = \tfrac{1}{2}(f + g - |f - g|), \text{ and } f \vee g = \tfrac{1}{2}(f + g + |f - g|),$$

so it is clearly enough to prove that $f \in A$ implies $|f| \in A$. We know that $f \in A$ implies $P \circ f \in A$ for all polynomials in one variable with $P(0) = 0$ (we never said that $A$ should contain the constants). Therefore it is enough to find a sequence of polynomials $(P_k)$ such that $P_k(0) = 0$ and $P_k(x) \to |x|$ uniformly on $[-1, 1]$. Indeed, if $f \in A$, then $|f| \leqslant c$ for some constant $c > 0$ and $cP_k(f/c)$ tends to $|f|$ uniformly on $K$. (This, of course, is a special case of the Weierstrass approximation theorem.)

One way to find such a sequence of polynomials is to use the binomial series with exponent $\frac{1}{2}$. We first have to extend the definition of the binomial coefficients (2.3) to non-integers:

$$\binom{z}{k} = \frac{z(z-1)\cdots(z-k-1)}{k!}, \qquad z \in \mathbf{C},\ k \in \mathbf{N}. \tag{5.1}$$

We know that, for complex $t$ such that $|1 - t^2| < 1$,

$$\begin{aligned} |t| = \sqrt{1 - (1 - t^2)} &= \sum_0^\infty \binom{\frac{1}{2}}{j}(-1)^j(1-t^2)^j \\ &= 1 - \sum_1^\infty \frac{1 \cdot 3 \cdot 5 \cdots (2j-3)}{2^j j!}(1-t^2)^j = 1 - \sum_1^\infty \frac{(2j-3)!!}{(2j)!!}(1-t^2)^j. \end{aligned}$$

Here $(-1)!! = 1$. Thus the series representation holds in particular for real $t$ satisfying $0 < t \leqslant 1$. But a closer analysis shows that it is valid also for $t = 0$, even uniformly for $-1 \leqslant t \leqslant 1$. For these values of $t$ the last series is majorized termwise by

$$\sum_1^\infty \frac{(2j-3)!!}{(2j)!!} = 1.$$

We can now define

$$P_k(t) = \sum_1^k \frac{(2j-3)!!}{(2j)!!}(1 - (1-t^2)^j).$$

We note that $P_k(0) = 0$ as desired and that $P_k(t) \to |t|$ uniformly for $t \in [-1, 1]$. The lemma is proved.

**Lemma 5.4.** *Let $L$ be a sublattice of $C(K)$, where $K$ is a compact space. Suppose that every function $f \in C(K)$ can be approximated by functions in $L$ at two arbitrary points, i.e., that to every $f \in C(K)$, every $\varepsilon > 0$, and every $s, t \in K$ there exists $g_{st} \in A$ such that $|g_{st}(x) - f(x)| < \varepsilon$ for $x = s, t$. Then $L$ is dense in $C(K)$.*

*Proof.* Fix an arbitrary function $f \in C(K)$ and define

$$U_{st} = \{x \in K;\, g_{st}(x) < f(x) + \varepsilon\} \text{ and } V_{st} = \{x \in K;\, g_{st}(x) > f(x) - \varepsilon\}.$$

For any fixed $s$ the family $(U_{st})_{t\in K}$ is an open covering of $K$. Since $K$ is compact, we can choose points $t_1, ..., t_m$ in $K$ such that $U_{st_1} \cup \cdots \cup U_{st_m} = K$. Then define $V_s = V_{st_1} \cap \cdots \cap V_{st_m}$ and $g_s = g_{st_1} \wedge \cdots \wedge g_{st_m} \in L$. We now have $g_s < f + \varepsilon$ in all of $K$, for if $x$ is an arbitrary point in $K$, then $x$ belongs to some $U_{st_j}$ and then $g_s(x) \leqslant g_{st_j}(x) < f(x) + \varepsilon$. Moreover, in $V_s$ we have $g_s > f - \varepsilon$, for $x \in V_s$ implies $g_{st_j} > f(x) - \varepsilon$ for all $j$, hence $g_s = \bigwedge_{j=1}^m g_{st_j} > f(x) - \varepsilon$.

Now $V_s$ is open as a finite intersection of open sets; $(V_s)_{s\in K}$ is an open covering of $K$. Choose finitely many points $s_1, ..., s_n$ in $K$ such that $V_{s_1} \cup \cdots \cup V_{s_n} = K$ and define $g = g_{s_1} \vee \cdots \vee g_{s_n} \in L$. In all of $K$ we now have $|f - g| < \varepsilon$, for $g_{s_j} < f + \varepsilon$ for all $j$, which implies $g = g_{s_1} \vee \cdots \vee g_{s_n} < f + \varepsilon$, and, on the other hand, if $x \in V_{s_j}$ then $g(x) \geqslant g_{s_j}(x) > f(x) - \varepsilon$, so that $g > f - \varepsilon$ in the union of the $V_{s_j}$, which is all of $K$.

**Theorem 5.5** (The Stone[7]–Weierstrass theorem). *Let $A$ be a subalgebra of $C(K)$, where $K$ is a compact topological space. Suppose that $A$ separates points in $K$, i.e., that for every pair of points $s, t \in K$ with $s \neq t$ there exists a function $g \in A$ such that $g(s) \neq g(t)$. Then either there exists a point $p \in K$ such that*

$$\overline{A} = C_p(K) = \{f \in C(K);\, f(p) = 0\},$$

*or else $\overline{A} = C(K)$.*

A special case is Weierstrass' approximation theorem: take $A = \bigcup \mathcal{P}_m\big|_I$, the algebra of all polynomials restricted to an interval $I$. We also immediately get the generalization to the case of a compact set $K$ in $\mathbf{R}^n$.

*Proof.* Suppose first that there exists, to any given $s \in K$ a function $h \in A$ such that $h(s) \neq 0$. Then, given any two points $s, t \in K$, $s \neq t$, there exists $g \in A$ such that $0 \neq g(s) \neq g(t) \neq 0$. Indeed, if it happened for instance that $g(s) = 0$, then we can replace $g$ by $g + \lambda h$, where $h$ is a function such that $h(s) \neq 0$ and $\lambda$ is sufficiently small. Then there exist, given arbitrary numbers $a, b \in \mathbf{R}$, real numbers $c, d$ such that

$$\begin{cases} cg(s) + dg^2(s) = a \\ cg(t) + dg^2(t) = b, \end{cases}$$

for the determinant of the system is $g(s)g(t)(g(t) - g(s)) \neq 0$. Since $cg + dg^2$ belongs to $A$ and $\overline{A}$ is a lattice according to Lemma 5.3, $\overline{A}$ satisfies the hypotheses of Lemma 5.4, and we conclude that $\overline{A} = C(K)$.

If, on the other hand, there exists a point $p \in K$ such that all elements $g$ of $A$ satisfy $g(p) = 0$, we can apply what we already proved to the algebra $A + \mathbf{R}$. To any given $f \in C_p(K)$ there exists a $g \in A$ and a $k \in \mathbf{R}$ such that $|f - g - k| < \varepsilon$. We note that $k$ must satisfy $|k| = |f(p) - g(p) - k| < \varepsilon$, so that $|f - g| \leqslant |f - g - k| + |k| < 2\varepsilon$. This proves that $A$ is dense in $C_p(K)$.

*Remark.* In the complex case things are different. We can let $A$ be the algebra of functions which are continuous on the closed unit disk $K$ in $\mathbf{C}$ and holomorphic in its interior. Then $A$ is closed and separates points in $K$ but is different from $C(K)$. However, if we consider algebras $A$ such that $f \in A$ implies $\overline{f} \in A$, then the results

[7]Marshall Harvey Stone, 1903 04 08–1989 01 09.

hold as in the real case; we just apply what we have proved to the algebras $\operatorname{Re} A$ and $\operatorname{Im} A$.

## 6. Chebyshev's theorems

Let us consider the space $C(I)$, where $I$ is a compact interval of the real line. It is a Banach space of infinite dimension, and its subspace $\mathcal{P}_m\big|_I$, the restrictions to $I$ of polynomials of degree at most $m$, is a finite-dimensional subspace. The distance from a function $f$ to this subspace is also the greatest lower bound for the errors when approximating the function $f$ by polynomials of degree at most $m$:

$$E_m(f) = d(f, \mathcal{P}_m) = \inf\big(\|f-p\|; p \in \mathcal{P}_m\big). \tag{6.1}$$

(To simplify notation I shall sometimes write $\mathcal{P}_m$ where I should have written $\mathcal{P}_m\big|_I$.) Since $\mathcal{P}_m$ is closed in $C(I)$ (see proof below), this distance is zero if and only if $f$ itself is a polynomial of degree at most $m$. The Weierstrass approximation theorem says that the error $E_m(f)$ tends to zero as $m \to \infty$, and we even have an idea of how fast it can tend to zero—see Corollary 2.4 and Theorem 3.3.

In this section we shall first prove that the infimum in (6.1) is attained, and then that it is attained at a unique point, although the norm we are using, the supremum norm $\|\cdot\| = \|\cdot\|_\infty$, is very far from being a Hilbert norm.

**Theorem 6.1** (Borel [1905]). *The infimum in (6.1) is attained for some polynomial in $\mathcal{P}_m$.*

*Proof.* We shall use the theorem that a continuous function on a compact set attains its infimum. So we first have to prove that the function $\varphi(p) = \|p-f\|$ is a continuous function of $p \in \mathcal{P}_m$. But what does that mean? We can use the topology induced from $C(I)$, or we can consider the topology from $\mathbf{R}^{m+1}$, using the coefficients $(a_0, a_1, \dots, a_m) \in \mathbf{R}^{m+1}$ as coordinates for $p(x) = \sum_0^m a_j x^j$. However, these two topologies agree; it is a well-known theorem that a finite-dimensional space can have only one Hausdorff (separated) vector space topology. In this particular case it is easy to argue as follows. We first note that

$$\|p\| = \sup\left|\sum_0^m a_j x^j\right| \leqslant \sup_{0\leqslant j\leqslant m}\sup_{x\in I}|x|^j \sum_0^m |a_j| = C\|a\|_1,$$

so that the norm of $p$ calculated in $C(I)$ is majorized by a constant times the $l^1$ norm of the coefficient vector. Therefore the supremum norm $\|p\|$ of $p$ is a continuous function on $\mathbf{R}^{m+1}$. Its infimum on the unit sphere of $\mathbf{R}^{m+1}$ is attained and therefore cannot be zero. This gives an inequality in the other direction, and shows that the two norms are actually equivalent. (It follows that $\mathcal{P}_m$ is closed in $C(I)$: $\mathcal{P}_m$ is complete for the norm in $\mathbf{R}^{m+1}$ and therefore for the norm induced by $C(I)$, hence closed in $C(I)$.)

It is easy to see that $\varphi(p) = \|p-f\|$ is a continuous function of $p$ for the topology induced by $C(I)$:

$$\varphi(p) - \varphi(q) = \|p-f\| - \|q-f\| \leqslant \|p-q\|.$$

If we interchange the roles of $p$ and $q$ we see that

$$|\varphi(p) - \varphi(q)| \leqslant \|p - q\|, \qquad p, q \in \mathcal{P}_m.$$

But from what we have remarked, this implies that $\varphi(p)$ depends continuously on $p$ if we use its coefficients as coordinates.

Now $\mathbf{R}^{m+1}$ is not compact, so we have to restrict the function $\varphi$ to a compact subset. Indeed $\varphi$ is large outside some compact set. To see this we remark that

$$\varphi(p) = \|p - f\| \geqslant \|p\| - \|f\| = \|p\| - \varphi(0) > \varphi(0) \text{ if } \|p\| > 2\varphi(0),$$

so that the infimum of the whole space $\mathcal{P}_m$ is the same as the infimum over the ball $K = \{p;\ \|p\| \leqslant 2\varphi(0)\}$. The set $K$ is closed and bounded in $\mathbf{R}^{m+1}$, and therefore $\varphi$ as a function of $(a_0, ..., a_m)$ attains its infimum there, and this infimum must be the same as the global infimum. This proves the theorem.

Thus there exists for every $m \in \mathbf{N}$ a best approximant $p_m$ to a given continuous function $f \in C(I)$. It is easy to see that for $m = 0$ this polynomial, a constant, is unique and equal to the mean values of $\sup f$ and $\inf f$. The best approximant is however unique for all $m$ as was proved by Chebyshev. We shall first introduce his criterion for a function to be a best approximant.

**Theorem 6.2** (Chebyshev[8] [1854]). *Let a function $f \in C(I)$ be given on $I = [a, b]$. Then a polynomial $p \in \mathcal{P}_m$ is a best approximant of $f$ in the supremum norm if and only if there exist at least $m+2$ points $x_j$ in $[a, b]$ with $a \leqslant x_1 < x_2 < \cdots < x_{m+2} \leqslant b$, such that the difference $r = f - p$ takes alternating values at the $x_j$ and its absolute value at each of the points is $\|r\|$: $(-1)^j r(x_j)$ is independent of $j$ and $|r(x_j)| = \|r\|$ for all $j$.*

The points $x_j$ are called *deviation points.*[9]

*Example 6.3.* Let $a = -4\pi$, $b = 4\pi$ and $f(x) = \sin x$. The zero polynomial satisfies the criterion in the theorem for $m = 6$, so it follows that the zero polynomial is a best approximant to $f$ in $\mathcal{P}_6$.

*Example 6.4.* Let $a = -1$, $b = 1$ and take $f(x) = x^m$, $m \in \mathbf{N}^*$. Let the best approximant (unique as we shall prove) in $\mathcal{P}_{m-1}$ be $p_{m-1}$. Then

$$x^m - p_{m-1}(x) = 2^{-m+1} \cos(m \arccos x) = 2^{-m+1} T_m(x),$$

where the last equation defines a polynomial $T_m$, called the *Chebyshev polynomial* of degree $m$. See Figure 4. The error is thus $2^{-m+1}$. The criterion is satisfied, as can be easily verified.

*Proof. Necessity.* Assume that $p$ is a best approximant of $f$. If $f$ itself is a polynomial of degree $m$, then $r$ is zero and the criterion is trivially satisfied for $p = f$. Otherwise we have $\|r\| > 0$ and we first note that we must have $\sup r = \|r\|$ and $\inf r = -\|r\|$.

[8]Pafnutij L′vovič Čebyšev, 1821–1894.

[9]In Russian the set of these $m + 2$ points is called *čebyševskij al′ternans.*

*Figure 4.* The Chebyshev polynomial $T_{40}$.

Indeed, if for instance we had $\sup r = \|r\| - 2\delta$ with $\delta > 0$, then we could find a better approximant $q = p - \delta$:

$$\|f - q\| = \|f - p + \delta\| = \max(\sup(r + \delta), -\inf(r + \delta)) = \|r\| - \delta < \|r\|.$$

Let us call a point $x$ such that $r(x) = \|r\|$ a *positive deviation point* and similarly a point $x$ such that $r(x) = -\|r\|$ a *negative deviation point.* Thus we have proved that there are at least one positive and at least one negative deviation point. For $m = 0$ the proof of the necessity is now complete. Otherwise we go on to find zeros $z_j$ of $r$ with the property that each of the intervals

$$[a, z_1], \ [z_1, z_2], \ldots, \ [z_{k-1}, z_k], \ [z_k, b]$$

contains only positive or only negative deviation points. To fix ideas we may assume that the first deviation point is positive. Then we choose $z_1$ as the largest zero such that $[a, z_1]$ contains only positive deviation points; then we choose $z_2$ as the largest zero such that $[z_1, z_2]$ contains at least one negative deviation point but no positive deviation point, provided there is such a zero; otherwise we stop and let $k = 1$. We go on; for compactness reasons the procedure must stop and there is a last zero $z_k$. There are now $k$ zeros $z_1, ..., z_k$ of $r$ and there are $k + 1$ deviation points $x_j$ with $x_1 < z_1 < \cdots < z_k < x_{k+1}$. If $k \leqslant m$ we find that the polynomial $p + q$, where $q(x) = \delta(x - z_1) \cdots (x - z_k)$ for a small $|\delta|$ is a better approximant than $p$ if we choose the sign of $\delta$ so that $q$ becomes positive in $[a, z_1[$ . Since this is against the hypothesis, we must have $k \geqslant m + 1$, and therefore there are at least $k + 1 \geqslant m + 2$ deviation points, as claimed.

*Sufficiency.* Let $p$ be a polynomial of degree at most $m$ which satisfies the criterion in the theorem, and assume that there is another polynomial $q \in \mathcal{P}_m$ which gives a better approximation, i.e., $\|f - q\| < \|f - p\|$, in particular

$$|f(x_j) - q(x_j)| \leqslant \|f - q\| < \|f - p\| = |f(x_j) - p(x_j)|$$

at every deviation point $x_j$ for $p$. This implies that the sign of

$$q(x_j) - p(x_j) = (f(x_j) - p(x_j)) - (f(x_j) - q(x_j))$$

is the same as the sign of $f(x_j) - p(x_j)$, so that $p - q$ changes its sign at least $m + 1$ times in $[a, b]$; it follows that $p - q$ has at least $m + 1$ zeros. Since $p - q$ is not identically zero and of degree at most $m$, this is impossible. Therefore there are no better approximants than $p$.

A variant of the last argument gives an even better result: the uniqueness of the best approximant:

**Theorem 6.5.** *For every $f \in C(I)$ and every $m \in \mathbf{N}$ there exists a unique polynomial $p_m$ in $\mathcal{P}_m$ such that*

$$\|p_m - f\| = \inf\bigl(\|p - f\|; p \in \mathcal{P}_m\bigr).$$

*Proof.* We argue as in the proof of the sufficiency in the last proof. Assume that $p$ and $q$ are both best approximants. Then also their mean value $\frac{1}{2}p + \frac{1}{2}q$ is a best approximant, for

$$E_m \leqslant \|f - (\tfrac{1}{2}p + \tfrac{1}{2}q)\| \leqslant \tfrac{1}{2}\|f - p\| + \tfrac{1}{2}\|f - q\| = E_m,$$

where $E_m = d(f, \mathcal{P}_m) = \|f - p\|$. By Theorem 6.2 there must exist $m + 2$ deviation points $x_j$ for $\frac{1}{2}p + \frac{1}{2}q$, which means that

$$E_m = |f(x_j) - (\tfrac{1}{2}p(x_j) + \tfrac{1}{2}q(x_j)| \leqslant \tfrac{1}{2}|f(x_j) - p(x_j)| + \tfrac{1}{2}|f(x_j) - q(x_j)| \leqslant E_m.$$

It follows from this not only that the $x_j$ are deviation points also for $p$ and $q$:

$$|f(x_j) - p(x_j)| = |f(x_j) - q(x_j)| = E_m,$$

but even that the differences $f(x_j) - p(x_j)$ and $f(x_j) - q(x_j)$ must have the same sign. Hence the differences are equal: $f(x_j) - p(x_j) = f(x_j) - q(x_j)$ for $j = 1, ..., m + 2$. The polynomial $p - q$ has $m + 2$ zeros and is of degree at most $m$. Hence it vanishes identically: $p = q$ and the best approximant is unique.

There is a variant of Theorem 6.2 for periodic functions.

**Theorem 6.6.** *Let a continuous function $f$ be defined on the real axis with period $2\pi$. Then a trigonometric polynomial $T$ of order $m$ is a best approximant of $f$ in the supremum norm if and only if there exist at least $2m+2$ points $t_j$ in an interval $[a, a+2\pi[$ with $a \leqslant t_1 < t_2 < \cdots < t_{2m+2} < a+2\pi$, such that the difference $r = f-T$ takes alternating values at the $t_j$ and its absolute value at each of the points is $\|r\|$: $(-1)^j r(t_j)$ is independent of $j$ and $|r(t_j)| = \|r\|$ for all $j$.*

The proof of this theorem is like that of Theorem 6.2. The polynomial $q$ in the proof of the necessity has to be replaced by

$$q(t) = \delta \sin\frac{t-z_1}{2}\sin\frac{t-z_2}{2}\cdots\sin\frac{t-z_k}{2}.$$

*Definition 6.7.* Let a metric space $M$ be given. A set of $m+1$ functions $\varphi_0, ..., \varphi_m \in C(M)$ is called a *Chebyshev system* if $M$ contains at least $m+1$ points and if a linear combination $\sum c_j\varphi_j$ can have $m+1$ different zeros only if all coefficients $c_j$ vanish.

*Examples.* The system $\varphi_j(z) = z^j$, $j = 0, ...., m$ is a Chebyshev system on $\mathbf{C}$, as well as on every subset of $\mathbf{C}$ containing at least $m+1$ points. The system $1, e^{iz}, e^{-iz}, ..., e^{imz}, e^{-imz}$ as well as the system $1, \cos z, \sin z, ..., \cos mz, \sin mz$ are Chebyshev on the strip $a \leqslant \operatorname{Re} z < a + 2\pi$ for any $a \in \mathbf{R}$, and of course also on any subset of the strip containing at least $2m+1$ points, in particular on the interval $[a, a+2\pi[$. However, it is not Chebyshev on the interval $[0, 2\pi]$, since $\sin mt$ vanishes at $2m+1$ points $t = k\pi/m$, $k = 0, ..., 2m$. (It is obvious that we should count in $\mathbf{R}$ modulo $2\pi$.) It follows that the system $1, \cos t, \cos 2t, ..., \cos mt$ is Chebyshev on $[0, \pi[$ and the system $\sin t, \sin 2t, ..., \sin mt$ is Chebyshev on $]0, \pi[$.

*Example.* The functions $\varphi_0(x) = x^2 - x$, $\varphi_1(x) = x^2 + x$, $\varphi_2(x) = x^2 + 1$ form a Chebyshev system on the real axis. Any two of these three functions do not form a Chebyshev system.

There is a criterion for being Chebyshev in terms of determinants:

**Exercise 6.8.** Let a metric space $M$ be given with at least $m+1$ points. Then $\varphi_0, ..., \varphi_m \in C(M)$ is a Chebyshev system if and only if the determinant $\det(\varphi_j(x_k))$, $j, k = 0, ..., m$, is nonzero for every choice of different points $x_0, ..., x_m$ in $M$.

**Exercise 6.9.** Let a Chebyshev system $\varphi_0, ..., \varphi_m$ on a metric space $M$ be given and let $x_0, ..., x_m$ be different points. Then for any real numbers $y_j$, $j = 0, ..., m$, there exists a linear combination $P$ of the $\varphi_k$ such that $P(x_j) = y_j$.

The importance of the notion of Chebyshev system for unique best approximants is made clear by the following theorem.

**Theorem 6.10** (Haar[10] [1918], Kolmogorov[11] [1948]). *Let $K$ be a compact subset of $\mathbf{R}^n$ containing at least $m+1$ points and let $\varphi_0, ..., \varphi_m$ be $m+1$ real-valued or complex-valued continuous functions on $K$. Then there is a unique best approximant in the supremum norm among the linear combinations of $\varphi_0, ..., \varphi_m$ for an arbitrary continuous function on $K$ if and only if $\varphi_0, ..., \varphi_m$ is a Chebyshev system.*

[10] Alfred Haar, 1885–1933.

[11] Andrej Nikolaevič Kolmogorov, 1903–1987.

I shall not give the proof here. Let me mention that it uses the exercises 6.8 and 6.9 and the arguments developed in the proof of Theorems 6.2 and 6.5.

Nice as this theory may be, it is of limited interest in several variables, for Mairhuber [1956] has proved that if there exists a Chebyshev system with $m \geqslant 1$ on a compact space $K$, then $K$ is homeomorphic to a circle or a subset of a circle. So approximation in several variables is different, and the reader is asked to ponder the following exercise:

**Exercise 6.11.** Let $K = [0,1]^2 \subset \mathbf{R}^2$ and define $\varphi_0(x) = 1$, $\varphi_1(x) = x_1$, $\varphi_2(x) = x_2$. Then there exist functions in $C(K)$ which do not admit a unique best approximant among the linear combinations of $\varphi_0, \varphi_1, \varphi_2$.

## 7. Approximation by polynomials and trigonometric polynomials

The purpose of this section is twofold: to introduce results on trigonometric polynomials and approximation by them and, secondly, to prove best possible results—it turns out that the Bernstein polynomials do not give the best possible approximation.

Let us review first some elementary results from the theory of Fourier series. We shall work with periodic functions; it is convenient to let the period be $2\pi$. Let us denote by $C_{2\pi}(\mathbf{R})$ the space of continuous functions on the real line which are periodic with period $2\pi$. It is the same thing as the continuous functions on the quotient space $\mathbf{R}/2\pi = \mathbf{R} \bmod 2\pi$, thus $C_{2\pi}(\mathbf{R}) = C(\mathbf{R}/2\pi)$.

A function $f$ in $C_{2\pi}(\mathbf{R})$ has a Fourier series

$$f(x) \sim \tfrac{1}{2}a_0 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx), \tag{7.1}$$

where the Fourier coefficients $a_k$, $k \in \mathbf{N}$, and $b_k$, $k \in \mathbf{N}^*$, are defined by

$$a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx, \qquad b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\,dx. \tag{7.2}$$

We also know about two important kernels, the *Dirichlet*[12] *kernel* $D_m$ and the *Fejér*[13] *kernel* $F_m$. To wit:

$$D_m(x) = \tfrac{1}{2} + \cos x + \cos 2x + \cdots + \cos mx = \frac{\sin(m+\frac{1}{2})x}{2\sin\frac{1}{2}x}, \tag{7.3}$$

$$F_m(x) = \frac{1}{m}\sum_{0}^{m-1} D_k(x) = \frac{1}{2m}\left(\frac{\sin\frac{m}{2}x}{\sin\frac{1}{2}x}\right)^2. \tag{7.4}$$

The function $D_m$ is not positive, and its total mass $\int_0^{2\pi}|D_m(x)|dx$ tends to infinity like $\log m$. The kernel $F_m$, on the other hand, is positive.

[12] Peter Gustav Lejeune Dirichlet, 1805–1859.

[13] Lipót Fejér, 1880–1959, my great-grand-advisor.

We define the convolution product $f * g$ of two functions in $C_{2\pi}(\mathbf{R})$ as

$$(f * g)(x) = \frac{1}{\pi} \int_0^{2\pi} f(y) g(x-y)\, dy, \qquad x \in \mathbf{R}. \tag{7.5}$$

With this normalization we have $f * 1 = a_0 =$ twice the mean value of $f$. Also $(f * \cos)(0) = a_1$, and so on for the higher Fourier coefficients. More generally,

$$\begin{aligned}(f * (t \mapsto \cos kt))(x) &= \frac{1}{\pi} \int_0^{2\pi} f(y) \cos k(x-y)\, dy \\ &= \frac{1}{\pi} \int_0^{2\pi} f(y) \cos kx \cos ky\, dy + \frac{1}{\pi} \int_0^{2\pi} f(y) \sin kx \sin ky\, dy \\ &= a_k \cos kx + b_k \sin kx.\end{aligned}$$

Summing over $k = 0, ..., m$, we see that $f * D_m$ is nothing but the partial sum of the Fourier series.

We also know from some course on Fourier analysis that the convolution product $f * D_m$ converges uniformly to $f$ if $f$ is of class $C^1$, but that there exist continuous functions such that this partial sum does not converge at a given point. On the other hand, the convolution $f * F_m$ always converges uniformly to $f$ if $f$ is continuous. This yields the Weierstrass approximation theorem for trigonometric polynomials:

**Theorem 7.1.** *Let $f \in C_{2\pi}(\mathbf{R})$ and let $\varepsilon$ be a positive number. Then there exists a trigonometric polynomial $T$ such that $|T(x) - f(x)| < \varepsilon$ for all $x \in \mathbf{R}$.*

*Proof.* The Fejér kernel does the job: $f * F_m$, $m \in \mathbf{N}^*$, are trigonometric polynomials, and they converge uniformly to $f$ for every $f \in C_{2\pi}(\mathbf{R})$ as $m \to \infty$.

We shall now try to improve the results of Corollary 2.4, which gave us polynomials converging to $f$ with an error $\frac{5}{4}\omega(1/\sqrt{m})$.

We shall sum Fourier series in a more general way. Let us define

$$\sigma_m(x) = \tfrac{1}{2} + \sum_{k=1}^{\infty} \rho_{mk} \cos kx, \qquad x \in \mathbf{R},\ m \in \mathbf{N}^*,$$

where the $\rho_{mk}$, $m, k \in \mathbf{N}^*$, are real numbers and where for each fixed $m$, $\rho_{mk} = 0$ for large $k$. So the Dirichlet kernel is the special case when $\rho_{mk} = 1$ for $k = 1, ..., m$ and zero otherwise; the Fejér kernel the special case when $\rho_{mk} = 1 - k/m$ for $k = 1, ..., m-1$ and zero otherwise. The convolution

$$(f * \sigma_m)(x) = \frac{1}{\pi} \int_0^{2\pi} f(y) \sigma_m(x-y) dy = \tfrac{1}{2} a_0 + \sum_{k=1}^{\infty} \rho_{mk}(a_k \cos kx + b_k \sin kx) \tag{7.6}$$

is well defined for any integrable periodic function, in particular for any function in $C_{2\pi}(\mathbf{R})$, and is a trigonometric polynomial of order at most equal to the largest $k$ such that $\rho_{mk}$ is nonzero.

We shall define a kernel, known as the *Jackson*[14] *kernel* as follows:

$$(7.7)\qquad J_m(x) = \frac{6m}{2m^2+1}F_m(t)^2 = \frac{6m}{2m^2+1}\left[\tfrac{1}{2} + \sum_{j=1}^{m-1}(1-j/m)\cos jx\right]^2$$
$$= \tfrac{1}{2}j_0 + \sum_{k=1}^{2m-2} j_k \cos kx.$$

It is clear that $J_m$ is a positive trigonometric polynomial of order $2m-2$. It is also obvious that $c_m F_m^2$ should be a good kernel in the sense of Theorem 4.1 (adapted to the case of periodic functions) for some choice of $c_m > 0$. So the only thing we need to check is that $c_m = 6m/(2m^2+1)$ gives the right mean value. This is easy:

$$j_0 = \frac{1}{\pi}\int_0^{2\pi} J_m(t)dt = \frac{6m}{2m^2+1}\left[\tfrac{1}{2} + \sum_1^{m-1}(1-k/m)^2\right]$$
$$= \frac{6m}{2m^2+1}\left(\tfrac{1}{2} + \frac{(m-1)m(2m-1)}{6m^2}\right) = 1,$$

where the second equality follows from the orthogonality of the cosine functions on $[0, 2\pi]$.

**Exercise 7.2.** Calculate $j_1 = \rho_{2m-2,1}$ for the Jackson kernel.

**Lemma 7.3.** *The first moment of the Jackson kernel can be estimated as follows:*

$$\frac{1}{\pi}\int_0^{\pi} tJ_m(t)dt \leqslant \frac{3\pi}{4m}, \qquad m \in \mathbf{N}^*.$$

*Proof.* The integral in question is equal to

$$\frac{3}{2\pi m(2m^2+1)}\left[\int_0^{\pi/m} + \int_{\pi/m}^{\pi}\right] t\left(\frac{\sin\frac{mt}{2}}{\sin\frac{t}{2}}\right)^4 dt$$
$$\leqslant \frac{3}{4\pi m^3}\left[\int_0^{\pi/m} tm^4 dt + \int_{\pi/m}^{\pi} t\frac{\pi^4}{t^4}dt\right] \leqslant \frac{3\pi}{4m}.$$

**Exercise 7.4.** Prove that the first moment of the Fejér kernel satisfies

$$c\frac{\log m}{m} \leqslant \int_0^{\pi} tF_m(t)dt \leqslant C\frac{\log m}{m}, \qquad m \geqslant 2,$$

for some positive constants $c, C$.

[14]Dunham Jackson, 1888–1946.

Let us now estimate how far the trigonometric polynomial $f * J_\nu$ is from $f$ in supremum norm. We note that

$$f(x) - (f * J_\nu)(x) = \frac{1}{\pi}\int_0^{2\pi} \big[f(x)J_\nu(t) - f(x-t)J_\nu(t)\big]\,dt,$$

so that

$$\begin{aligned}(7.8)\quad |(f - f * J_\nu)(x)| &\leqslant \frac{1}{\pi}\int_0^{\pi} \big[|f(x+t) - f(x)| + |f(x-t) - f(x)|\big]\,J_\nu(t)\,dt \\ &\leqslant \frac{2}{\pi}\int_0^{\pi} \omega(t)J_\nu(t)\,dt \leqslant \omega(1/\nu)\frac{2}{\pi}\int_0^{\pi}(\nu t + 1)J_\nu(t)\,dt \\ &\leqslant \omega(1/\nu)\left(2\nu\frac{3\pi}{4\nu} + 1\right) = \omega(1/\nu)\left(\frac{3\pi}{2} + 1\right) \leqslant 6\omega(1/\nu).\end{aligned}$$

Now $f * J_\nu$ is of order $2\nu - 2$. Given any $m$, we define $\nu = m/2 + 1$ in case $m$ is even, and $\nu = (m-1)/2 + 1$ in case $m$ is odd. Then $f * J_\nu$ is of order at most $2\nu - 2 \leqslant m$ and $\omega(1/\nu) \leqslant 2\omega\big(\frac{1}{m+1}\big)$ in all cases, so that

$$\|f - f * J_\nu\| \leqslant 6\omega(1/\nu) \leqslant 12\omega\big(\tfrac{1}{m+1}\big).$$

**Theorem 7.5.** (P. P. Korovkin [1959]). *Suppose that we have real numbers $\rho_{mk}$, $k, m \in \mathbf{N}^*$, satisfying*

$$(7.9)\qquad \sigma_m(x) = \tfrac{1}{2} + \sum_{k=1}^{m} \rho_{mk}\cos kx \geqslant 0, \qquad x \in \mathbf{R},$$

*and*

$$(7.10)\qquad \rho_{m1} \to 1 \text{ as } m \to \infty.$$

*Then the convolution $f * \sigma_m$ converges uniformly to $f$ for any $f \in C_{2\pi}(\mathbf{R})$.*

This theorem is a consequence of the following more precise result:

**Theorem 7.6.** (P. P. Korovkin [1959]). *If the $\rho_{mk}$ are chosen so that $\sigma_m \geqslant 0$, then*

$$(7.11)\qquad \|f * \sigma_m - f\|_\infty \leqslant \omega(\delta)\left(1 + \frac{\pi}{\sqrt{2}\delta}\sqrt{1 - \rho_{m1}}\right), \qquad \delta > 0,\ m \in \mathbf{N}^*,$$

*where $\omega$ denotes the modulus of continuity of $f$.*

The remarkable thing here is that conditions (7.10) and (7.11) involve only $\rho_{m1}$, whereas the $\rho_{mk}$ for $k \geqslant 2$ do not appear—they play a role of course in the positivity condition.

**Lemma 7.7.** *The first moment of the kernel $\sigma_m$ can be estimated as follows:*

$$\frac{1}{\pi}\int_0^\pi t\sigma_m(t)dt \leqslant \frac{\pi}{2\sqrt{2}}\sqrt{1-\rho_{m1}}, \qquad m \in \mathbf{N}^*.$$

*Proof.* We use first the estimate $t/\pi \leqslant \sin(t/2)$, and then the Hölder inequality:

$$\begin{aligned}\frac{1}{\pi}\int_0^\pi t\sigma_m(t)dt &\leqslant \int_0^\pi \sin\frac{t}{2}\sigma_m(t)dt = \int_0^\pi \sin\frac{t}{2}\sqrt{\sigma_m(t)}\sqrt{\sigma_m(t)}dt \\ &\leqslant \sqrt{\int_0^\pi \sin^2\frac{t}{2}\sigma_m(t)dt}\sqrt{\int_0^\pi \sigma_m(t)dt}.\end{aligned}$$

Here the expression under the first radical sign is

$$\int_0^\pi (\tfrac{1}{2}-\tfrac{1}{2}\cos t)\sigma_m(t)dt = \frac{\pi}{4}-\frac{\pi}{4}\rho_{m1},$$

and the expression under the second radical sign is $\pi/2$. This proves the lemma.

*Proof of Theorem 7.6.* The proof is just like the one for the Jackson kernel: we first use the inequality $\omega(t)=\omega((t/\delta)\delta)\leqslant(1+t/\delta)\omega(\delta)$, where $\delta$ is independent of $m$, and then Lemma 7.7 to estimate the first moment:

$$\begin{aligned}|f(x)-(f*\sigma_m)(x)| &\leqslant \frac{2}{\pi}\int_0^\pi \omega(t)\sigma_m(t)dt \leqslant \omega(\delta)\frac{2}{\pi}\int_0^\pi (1+t/\delta)\sigma_m(t)dt \\ &= \omega(\delta)\left(\frac{2}{\pi}\int_0^\pi \sigma_m(t)dt + \frac{2}{\pi\delta}\int_0^\pi t\sigma_m(t)dt\right) \leqslant \omega(\delta)\left(1+2\frac{\pi}{2\sqrt{2}\delta}\sqrt{1-\rho_{m1}}\right).\end{aligned}$$

**Corollary 7.8.** *For the Fejér kernel we have*

$$\|f*F_m-f\|_\infty \leqslant \left(1+\frac{\pi}{\sqrt{2}}\right)\omega(1/\sqrt{m}), \qquad m\in\mathbf{N}^*.$$

*Proof.* We know that the Fejér kernel is positive and that it satisfies $\rho_{m1}=1-1/m$. We choose $\delta=1/\sqrt{m}$ in Theorem 7.6 to conclude.

**Theorem 7.9** (Jackson's first theorem). *For any $f\in C_{2\pi}(\mathbf{R})$ and every $m\in\mathbf{N}^*$ there exists a trigonometric polynomial $T$ of order at most $m$ such that*

$$\|f-T\|_\infty \leqslant 6\omega\big(\tfrac{1}{m+1}\big).$$

*Proof.* Following Korovkin we shall define polynomials $\sigma_m \geqslant 0$ with

$$\rho_{m1}=\cos\frac{\pi}{m+2} \geqslant 1-\frac{\pi^2}{2(m+2)^2},$$

so that

$$\sqrt{1-\rho_{m1}} \leqslant \frac{\pi}{\sqrt{2}(m+2)},$$

which is a better estimate than that for the Fejér kernel. With this estimate we may choose $\delta = 1/(m+1)$ in (7.11) and obtain

$$\begin{aligned} \omega\left(\tfrac{1}{m+1}\right)\left(1+\frac{\pi(m+1)}{\sqrt{2}}\sqrt{1-\rho_{m1}}\right) &\leqslant \omega\left(\tfrac{1}{m+1}\right)\left(1+\frac{\pi(m+1)}{\sqrt{2}}\frac{\pi}{\sqrt{2}(m+2)}\right) \\ &\leqslant \omega\left(\tfrac{1}{m+1}\right)\left(1+\frac{\pi^2}{2}\right) \leqslant 6\omega\left(\tfrac{1}{m+1}\right), \end{aligned}$$

the estimate in Theorem 7.9. (The Jackson kernel has $\rho_{2m-2,1} = 1 - \frac{3}{2m^2}$, thus $\sqrt{1-\rho_{2m-2,1}} = \frac{\sqrt{3/2}}{m}$, almost as good as for the Korovkin kernel.)

So we shall find a positive kernel with this wonderful property. Define

$$\sigma_m(t) = A_m |a_1 + a_2 z + a_3 z^2 + \cdots + a_{m+1} z^m|^2,$$

where

$$z = \cos t + i \sin t,\ A_m = \frac{1}{2(a_1^2 + \cdots + a_{m+1}^2)},\ \text{and } a_k = \sin\frac{k\pi}{m+2}.$$

We find that

(7.12)

$$\rho_{m1} = \frac{a_1a_2 + a_2a_3 + \cdots + a_m a_{m+1}}{a_1^2 + a_2^2 + \cdots + a_{m+1}^2} = \frac{\sum_1^m \sin\frac{k\pi}{m+2}\sin\frac{k+1}{m+2}}{\sin^2\frac{\pi}{m+2} + \sin^2\frac{2\pi}{m+2} + \cdots + \sin^2\frac{(m+1)\pi}{m+2}}.$$

This expression can be calculated, using the formula

$$\cos\frac{\pi}{m+2}\sin^2\frac{k\pi}{m+2} = \tfrac{1}{2}\sin\frac{(k-1)\pi}{m+2}\sin\frac{k\pi}{m+2} + \tfrac{1}{2}\sin\frac{k\pi}{m+2}\sin\frac{(k+1)\pi}{m+2}.$$

Summing over $k = 1, ..., m+1$ we get a left-hand side equal to

$$\cos\frac{\pi}{m+2}\sum_1^{m+1}\sin^2\frac{k\pi}{m+2},$$

and a right-hand side equal to the numerator in the last expression in (7.12). This proves that $\rho_{m1} = \cos\frac{\pi}{m+2}$.

**Exercise 7.10.** We have seen that for the Jackson kernel and the Korovkin kernel we have $\sqrt{1-\rho_{m1}} = O(1/m)$, $m \to \infty$, which yields a good approximation. Show that we cannot have $\sqrt{1-\rho_{m1}} = o(1/m)$, $m \to \infty$, as long as $\sigma_m \geqslant 0$.

**Exercise 7.11.** Prove that for the Jackson kernel we have

$$\lim_{m\to\infty} m^2(J_m * f - f) = \frac{3}{2}f''$$

for functions of class $C^2$. Cf. Theorem 3.3.

It is known that one has

$$E_m(f) \leqslant \frac{\lambda}{2(m+1)}, \qquad m \in \mathbf{N},\ f \in C([0,1]), \tag{7.13}$$

for functions with Lipschitz constant $\lambda$ on $[0,1]$. It is easy to see that this estimate cannot be improved. To do so we can use a simple sawtooth function: define $f$ to take the values $(-1)^k$ at the points $x = k/(m+1)$, $k = 0, ..., m+1$, and to be affine in between these points. Then its Lipschitz constant is $\lambda = 2(m+1)$, and by the Chebyshev criterion (Theorem 6.2), the best approximant in $\mathcal{P}_m$ is zero. Therefore $E_m(f) = \|f\| = 1$. We see that in any estimate $E_m(f) \leqslant C\lambda/(m+1)$ we must have $C \geqslant \frac{1}{2}$. For the interval $[-1,1]$ the estimate is $E_m(f) \leqslant \lambda/(m+1)$ and it is also best possible.

Similarly for the periodic case it is known that

$$E_m(f) \leqslant \frac{\pi\lambda}{2(m+1)} \qquad m \in \mathbf{N},\ f \in C_{2\pi}(\mathbf{R}), \tag{7.14}$$

if $\lambda$ is the Lipschitz constant of $f$. Here we define a function which takes the value $\pm 1$ at the points $k\pi/(m+1)$, $k = 0, ..., 2m+2$, and is affine in between. There are $2m+2$ deviation points in a period $[0, 2\pi[$; note that $0 = 2\pi$ in the quotient space. This function shows that in any estimate $E_m(f) \leqslant C/(m+1)$ we must have $C \geqslant \pi/2$.

For general continuous functions the best estimate is

$$E_m(f) \leqslant \omega(\pi/(m+1)) \qquad m \in \mathbf{N}, \tag{7.15}$$

a result due to Kornejčuk [1962]. It can be shown using piecewise affine functions that this estimate is also best possible among those with $E_m(f) \leqslant C\omega(1/(m+1))$. However these piecewise affine functions are a little more complicated to describe than those we used in the Lipschitz case.

A more careful analysis yields an improvement of Theorem 7.9:

**Theorem 7.12.** *There exists a constant $C$ such that*

$$E_m(f) \leqslant C\omega_2\big(\tfrac{1}{m+1}\big), \qquad m \in \mathbf{N},\ f \in C_{2\pi}(\mathbf{R}),$$

*where*

$$\omega_2(\delta) = \sup_{x,t \in \mathbf{R}} \big[|f(x+2t) - 2f(x+t) + f(x)|; |t| \leqslant \delta\big]$$

*is the second modulus of continuity.*

*Proof.* We estimate like in (7.8) but a little more carefully, using the fact that $J_\nu$ is even:

$$\begin{aligned} f(x) - (f * J_\nu)(x) &= \frac{1}{\pi}\int_0^{2\pi} \big[f(x)J_\nu(t) - f(x-t)J_\nu(t)\big]\,dt \\ &= \frac{1}{\pi}\int_0^{\pi} (2f(x) - f(x-t) - f(x+t))J_\nu(t)\,dt, \end{aligned}$$

so that

$$|(f - f * J_\nu)(x)| \leqslant \frac{2}{\pi} \int_0^\pi \omega_2(t) J_\nu(t) dt.$$

We then use the fact that $\omega_2(t) = \omega_2((m+1)t \cdot (1/(m+1))) \leqslant (1+(m+1)t)^2 \omega_2\big(\frac{1}{m+1}\big)$ to deduce that

$$\|f - f * J_\nu\| \leqslant \omega_2\big(\tfrac{1}{m+1}\big) \int_0^\pi (1 + (m+1)t)^2 J_\nu(t) dt,$$

and the last integral is bounded; the second moment $\int_0^\pi t^2 J_\nu(t) dt$ can be estimated as in Lemma 7.7 and one can prove that it does not exceed a constant times $(m+1)^{-2}$. This proves the theorem.

There are higher moduli of continuity. We first define the differences

$$\Delta_t^1 f(x) = f(x+t) - f(x),$$
$$\Delta_t^{p+1} f(x) = \Delta_t^1 \Delta_t^p f(x), \qquad p \in \mathbf{N}^*.$$

Then the *pth modulus of continuity* is

$$\omega_p(f; \delta) = \sup_{\substack{x,t \in \mathbf{R} \\ |t| \leqslant \delta}} |\Delta_t^p f(x)|.$$

One can prove a generalization of Theorem 7.12:

$$E_m(f) \leqslant C_p \omega_p\big(f; \tfrac{1}{m+1}\big). \tag{7.16}$$

The sequence of errors $E_m(f)$ tends to zero if $f$ is continuous as we know. But how slowly? In fact arbitrarily slowly as the next theorem shows.

**Theorem 7.13.** *Given any decreasing sequence $(\varepsilon_m)_{m\in\mathbf{N}}$ of positive numbers tending to zero there exists a function $f \in C([-1,1])$ such that $E_m(f) \geqslant \varepsilon_m$, $m \in \mathbf{N}$.*

*Proof.* Put $\alpha_k = \varepsilon_{k-1} - \varepsilon_k \geqslant 0$ and define $f = \sum_0^\infty \alpha_k T_{3^k}$, where $T_m$ denotes the Chebyshev polynomial of degree $m$. Since $\|T_m\| \leqslant 1$ and $\sum \alpha_k < +\infty$, $f$ is continuous. We claim that $P = \sum_0^m \alpha_k T_{3^k}$ is the best approximant of degree at most $3^m$. Consider the error function $r = f - P = \sum_{m+1}^\infty \alpha_k T_{3^k}$. If we can show that it has at least $3^m + 2$ deviation points it will follow by Chebyshev's criterion that $P$ is a best approximant. Put $x_j = \cos(j\pi/3^{m+1})$, $j = 0, ..., 3^{m+1}$. For $k \geqslant m+1$ we have $T_{3^k}(x_j) = \cos(3^k j\pi/3^{m+1}) = (-1)^j$ (here it is important that $3^k j/3^{m+1}$ has the same parity as $j$). Thus $T_{3^k}(x_j)$ does not depend on $k$ so

$$r(x_j) = (-1)^j \sum_{k=m+1}^\infty \alpha_k = (-1)^j \varepsilon_m,$$

and we have

$$\varepsilon_m = |r(x_j)| \leqslant \|r\| \leqslant \sum_{m+1}^\infty \alpha_k = \varepsilon_m,$$

showing that the $x_j$ are in fact deviation points, and that there are at least $3^{m+1}+1 \geqslant 3^m+2$ of them. So $P$ is indeed a best approximant and $E_m(f) \geqslant E_{3^m}(f) = \|r\| = \varepsilon_m$.

## 8. The nonexistence of a continuous linear projection

Let us denote by $\pi_m(f)$ the polynomial in $\mathcal{P}_m$ closest to $f$. In view of Theorem 6.5 there is thus a well-defined operator $\pi_m: C(I) \to \mathcal{P}_m$. We may inquire as to its properties. Obviously it is idempotent, $\pi_m \circ \pi_m = \pi_m$; it is a projection onto $\mathcal{P}_m$. Moreover, it is positively homogeneous, i.e., it satisfies $\pi_m(\lambda f) = \lambda \pi_m(f)$ for every positive $\lambda$. It is easy to see by examples that $\pi_m(f+g) \neq \pi_m(f) + \pi_m(g)$ in general. So $\pi_m$ is certainly nonlinear. The Weierstrass theorem says that $\pi_m(f) \to f$ for the norm in $C(I)$ as $m \to \infty$. The operator is also continuous, as we shall prove now:

**Theorem 8.1.** *Let $\pi_m(f)$ denote the polynomial of degree at most $m \in \mathbf{N}$ which is closest to $f \in C(I)$ in the supremum norm. Then $\pi_m$ is a continuous operator; more precisely, for every $m \in \mathbf{N}$ and every $f \in C(I)$ there exists a positive number $C = C_{m,f}$ such that for all $g \in C(I)$,*

$$\|\pi_m(g) - \pi_m(f)\| \leqslant C\|g - f\|.$$

**Lemma 8.2.** *For every $m \in \mathbf{N}$ and every $f \in C(I)$ there exists a positive number $\gamma = \gamma_{m,f}$ such that for every polynomial $Q \in \mathcal{P}_m$ we have*

$$\|Q - f\| \geqslant \|\pi_m(f) - f\| + \gamma\|Q - \pi_m(f)\|.$$

Thus the function $Q \mapsto \|Q - f\|$ on $\mathcal{P}_m$, which attains its infimum at $Q = \pi_m(f)$, has a sharp minimum at that point: the value increases fast as we move away from $\pi_m(f)$.

*Proof.* If $f \in \mathcal{P}_m$, then $\pi_m(f) = f$ and the inequality holds with $\gamma = 1$. Now assume that $\|\pi_m(f) - f\| > 0$. According to Theorem 6.2 we have $m+2$ deviation points $x_j$ and signs $\theta_j = \pm 1$ such that, for every $j = 1, ..., m+2$, $\|\pi_m(f) - f\| = \theta_j(\pi_m(f) - f)(x_j)$. For any $Q \in \mathcal{P}_m \smallsetminus \{0\}$ we have $\max_j \theta_j Q(x_j) > 0$, for if we assume that the maximum is nonpositive we would get $\theta_j Q(x_j) \leqslant 0$ for all $j$, hence, since $\theta_{j+1} = -\theta_j$, that $Q$ has at least $m+1$ zeros counting multiplicities. In fact, we can prove inductively that $Q$ must have $k-1$ zeros in $[x_1, x_k]$. First, assume for definiteness that $Q(x_1) \leqslant 0$, $Q(x_2) \geqslant 0$ and so on. If $Q$ has a zero in $[x_1, x_2[$, we are done for $k = 2$. If not, then $x_2$ must be a zero. If $Q$ has a zero in $]x_2, x_3[$ we have two different zeros in $[x_1, x_3[$ ; if not, $Q$ is either positive or negative throughout $]x_2, x_3[$. In the first case $Q(x_3) = 0$; in the second, $x_2$ must be a double zero, so that, in either case, there are two zeros in $[x_1, x_3]$ counting multiplicities. It is now clear how to go on. Thus we have proved by contradiction that $\max \theta_j Q(x_j) > 0$. In view of the compactness we even have

$$\inf_{\|Q\|=1} \max_j \theta_j Q(x_j) = \gamma > 0. \tag{8.1}$$

The choice of points $x_j$ implies that

$$\begin{aligned}\|Q - f\| \geqslant \theta_j(Q(x_j) - f(x_j)) &= \theta_j(\pi_m(f) - f)(x_j) + \theta_j(Q - \pi_m(f))(x_j)\\ &= \|\pi_m(f) - f\| + \theta_j(Q - \pi_m(f))(x_j).\end{aligned}$$

We now take the maximum over all $j$:

$$\|Q-f\| \geqslant \|\pi_m(f)-f\| + \max_j \theta_j(Q-\pi_m(f))(x_j) \geqslant \|\pi_m(f)-f\| + \gamma\|Q-\pi_m(f)\|,$$

applying (8.1) in the last step to the normalized polynomial $(Q-\pi_m(f))/\|Q-\pi_m(f)\|$.

*Proof of Theorem 8.1.* Choose $\gamma$ according to Lemma 8.2 so that $\|Q-f\| \geqslant \|\pi_m(f)-f\| + \gamma\|Q-\pi_m(f)\|$ for all $Q \in \mathcal{P}_m$. In particular we may apply this inequality to $Q = \pi_m(g)$ and obtain

$$\begin{aligned}\|\pi_m(f)-f\| + \gamma\|\pi_m(g)-\pi_m(f)\| &\leqslant \|\pi_m(g)-f\| \leqslant \|\pi_m(g)-g\| + \|g-f\| \\ &\leqslant \|\pi_m(f)-g\| + \|g-f\| \leqslant \|\pi_m(f)-f\| + \|f-g\| + \|g-f\| \\ &= \|\pi_m(f)-f\| + 2\|g-f\|.\end{aligned}$$

Subtracting $\|\pi_m(f)-f\|$ from the first and the last expression above we see that $\gamma\|\pi_m(g)-\pi_m(f)\| \leqslant 2\|g-f\|$, proving the theorem with the constant $C = 2/\gamma$.

For trigonometric polynomials we have a similar projection $\tau_m\colon C_{2\pi}(\mathbf{R}) \to \mathcal{T}_m$. Theorem 8.1 has a counterpart for $\tau_m$ and even for general Chebyshev systems, since the characterization of best approximant, which was used in Lemma 8.2, holds also in these cases.

We now ask whether there exists an operator like $\pi_m$ or $\tau_m$ which in addition is linear. The answer is in the negative as is shown by the following result.

**Theorem 8.3** (Kharshiladze and Lozinskij[15]; Lozinskij [1948]). *Let there be given, for each $m \in \mathbf{N}$, a continuous linear projection $L_m$ of $C_{2\pi}(\mathbf{R})$ onto the space of trigonometric polynomials of order at most $m$. Then there exists an $f \in C_{2\pi}(\mathbf{R})$ for which $\{\|L_m(f)\|; m \in \mathbf{N}\}$ is unbounded.*

A key step in the proof is an estimate from below of the norm of any projection.

**Proposition 8.4.** *Let $L$ be a continuous linear projection of $C_{2\pi}(\mathbf{R})$ onto the subspace $\mathcal{T}_m$ of trigonometric polynomials of order at most $m$. Then $\|L\| \geqslant \|S_m\|$, where $S_m$ denotes the operator assigning to $f$ its partial Fourier series up to order $m$, i.e., $S_m(f) = f * D_m$ as defined by (7.3) and (7.5).*

*Proof.* Define the translation operator $T_a$ by

$$T_a(f)(x) = f(x-a), \qquad x \in \mathbf{R},\ a \in \mathbf{R}.$$

Then $T_a \circ L \circ T_{-a}$ is also a projection. We let $\widetilde{L}$ be the mean value of $T_a \circ L \circ T_{-a}$ over all translations:

$$\widetilde{L}(f)(x) = \frac{1}{2\pi}\int_0^{2\pi} (T_a \circ L \circ T_{-a})(f)(x)\,da, \qquad x \in \mathbf{R}.$$

We claim that $\widetilde{L} = S_m$. To prove this it is enough to prove that $\widetilde{L}(f_k) = S_m(f_k)$, where $f_k(x) = e^{ikx}$, $k \in \mathbf{Z}$, for the linear combinations of these functions are dense in

[15] Sergej Mihailovič Lozinskij, born 1914.

$C_{2\pi}(\mathbf{R})$. Now if $|k| \leqslant m$, $S_m f_k = f_k$ and $T_{-a} f_k \in \mathcal{T}_m$, so that $L \circ T_{-a}(f_k) = T_{-a}(f_k)$ and $(T_a \circ L \circ T_{-a})(f_k) = f_k$. Consequently the mean value $\widetilde{L}(f_k)$ is also equal to $f_k$. On the other hand, if $|k| > m$, then $S_m(f_k) = 0$. Now $T_{-a}(f_k) = e^{ika} f_k$ so that $(T_a \circ L \circ T_{-a})(f_k) = e^{ika} L(f_k)(x-a)$. But $L(f_k) \in \mathcal{T}_m$, so that, for any fixed $x$, $a \mapsto L(f_k)(x-a)$ is orthogonal to $a \mapsto e^{ika}$ for $|k| > m$. Thus also $\widetilde{L}(f_k) = 0$.

Now that we know that $\widetilde{L} = S_m$ we can easily finish the proof of the proposition:

$$\|\widetilde{L}(f)\| = \sup_x \left| \frac{1}{2\pi} \int_0^{2\pi} (T_a \circ L \circ T_{-a})(f)(x) dx \right| \leqslant \|T_a \circ L \circ T_{-a}\| \|f\| \leqslant \|L\| \|f\|,$$

thus $\|S_m(f)\| = \|\widetilde{L}(f)\| \leqslant \|L\| \|f\|$. Taking the supremum over all $f$ we get $\|S_m\| \leqslant \|L\|$ as claimed.

*Proof of Theorem 8.3.* We have $\|L_m\| \geqslant \|S_m\|$, and the latter norm is

$$\|S_m\| = \sup_{\|f\| \leqslant 1} \left| \frac{1}{\pi} \int_0^{2\pi} D_m(x) f(x) dx \right| = \frac{1}{\pi} \int_0^{2\pi} |D_m(x)| dx.$$

It is not difficult to prove that the latter quantity tends to infinity (in fact like a constant times $\log m$). So $\|L_m\|$ tends to infinity. We now apply the Banach–Steinhaus theorem, also known as the uniform boundedness principle: if $\{\|L_m(f); m \in \mathbf{N}\|\}$ is bounded for all $f$, then also $\{\|L_m\|; m \in \mathbf{N}\}$ is bounded.

One could also consider the functionals $f \mapsto L_m(f)(0)$ to get a function $f$ such that the set $\{L_m(f)(0); m \in \mathbf{N}\}$ is unbounded.

The Kharshiladze–Lozinskij theorem was formulated here for periodic functions but has counterparts for $C(I)$.

## 9. Approximation of functions of higher regularity

In this section we shall consider functions of higher regularity, meaning functions that are $k$ times continuously differentiable for some $k \in \mathbf{N}$. Let us write $C^k_{2\pi}(\mathbf{R})$ for the functions in $C^k(\mathbf{R})$ which are periodic with period $2\pi$.

The smallest error when approximating by trigonometric polynomials of order at most $m$ will be denoted by $E_m(f)$, thus

$$E_m(f) = \inf(\|P - f\|_\infty; P \in \mathcal{T}_m). \tag{9.1}$$

**Theorem 9.1.** *There exist a constant $C_0$ such that for all $k \in \mathbf{N}$, all $f \in C^k_{2\pi}(\mathbf{R})$ and all $m \in \mathbf{N}$,*

$$E_m(f) \leqslant \left(\frac{C_0}{m+1}\right)^k \|f^{(k)}\|.$$

*Proof.* In the proof we shall need to consider another error,

$$e_m(f) = \inf(\|P - f\|_\infty; P \in \mathcal{T}_m \text{ without constant term}), \tag{9.2}$$

i.e., when we approximate by trigonometric polynomials with mean value zero. Obviously $E_m(f) \leqslant e_m(f)$. We know from Jackson's theorem that $E_m(f) \leqslant C_m \|f'\|$ for every $f \in C^1_{2\pi}(\mathbf{R})$, where $C_m = C_0/(m+1)$. (We proved that one can take $C_0 = 6$; the best result is actually $C_0 = \pi/2$; cf. (7.13).) It is an important observation that the proof actually gives also that

$$e_m(f) \leqslant C_m \|f'\| \tag{9.3}$$

with the same constant if $f$ has mean value zero. This follows from the fact that the approximation is given by a convolution that reproduces the mean value, so that the approximant has mean value zero if the function has mean value zero, a fact that can be read off from (7.6); the mean value of $f * \sigma_m$ is $a_0/2$, the same as for $f$.

Given $f \in C^1_{2\pi}(\mathbf{R})$ without constant term we choose $p \in \mathcal{T}_m$ as the trigonometric polynomial which is the best approximant to $f'$ with mean value zero, thus $\|f' - p\| = e_m(f')$. Let $P \in \mathcal{T}_m$ be the primitive function to $p$ with mean value zero. Then, using (9.3),

$$e_m(f) = e_m(f - P) \leqslant C_m \|f' - p\| = C_m e_m(f').$$

Thus $e_m(f) \leqslant C_m e_m(f')$ if $f$ has mean value zero. Repeated application gives $e_m(f) \leqslant C_m^k e_m(f^{(k)})$ if $f$ is of class $C^k$, for $f'$ and all higher derivatives have mean value zero. Thus, given any $f \in C^k_{2\pi}(\mathbf{R})$, we conclude that

$$E_m(f) = E_m(f - a_0/2) \leqslant e_m(f - a_0/2) \leqslant C_m^k e_m(f^{(k)}) \leqslant C_m^k \|f^{(k)}\|.$$

This completes the proof.

**Theorem 9.2.** *For every $f \in C^1_{2\pi}(\mathbf{R})$ we have*

$$E_m(f) \leqslant \frac{C}{m+1} \omega\bigl(f', \tfrac{1}{m+1}\bigr), \qquad m \in \mathbf{N}.$$

*Proof.* From Theorem 7.12 we know that $E_m(f) \leqslant C\omega_2\bigl(\frac{1}{m+1}\bigr)$. It is easy to see that $\omega_2(f;\delta) \leqslant \delta\omega(f';\delta)$. Combining the two inequalities we get the desired conclusion.

Using (7.16) we can prove like in Theorem 9.2 that

$$E_m(f) \leqslant \frac{C_p}{(m+1)^p} \omega\bigl(f^{(p)}; \tfrac{1}{m+1}\bigr)$$

for every $f \in C^p_{2\pi}(\mathbf{R})$, $p \in \mathbf{N}$.

## 10. Inverse theorems

The theorems we have considered so far are traditionally called *direct theorems* (smoothness implies good approximation); we now come to the *inverse theorems*, i.e., those where we assume that we have a certain degree of approximation and conclude that the function possesses a certain smoothness. We shall consider here the case when the error $E_m(f)$ is $O(m^{-\alpha})$ for some real $\alpha$.

First a famous inequality:

**Theorem 10.1** (Bernstein's inequality). *Any trigonometric polynomial $T$ of order $m$ satisfies $\|T'\|_\infty \leqslant m\|T\|_\infty$.*

**Lemma 10.2.** *Let $P$ be a polynomial of degree $m-1$. Then*

$$\sup_{x\in[-1,1]} |P(x)| \leqslant m \sup_{x\in[-1,1]} |\sqrt{1-x^2}P(x)|. \tag{10.1}$$

*Proof.* Let $x_j = \cos((2j-1)\pi/2m)$, $j = 1, ..., m$, be the zeros of the Chebyshev polynomial $T_m$. Let $M$ denote the right-hand side of (10.1). If $x$ happens to belong to the interval $[x_m, x_1]$, it is easy to prove that $|P(x)| \leqslant M$:

$$\sqrt{1-x^2} \geqslant \sqrt{1-x_1^2} = \sqrt{1-\cos^2\frac{\pi}{2m}} = \sin\frac{\pi}{2m} \geqslant \frac{1}{m},$$

so that $|P(x)| \leqslant m\sqrt{1-x^2}|P(x)|$ for these values of $x$.

The case when $x$ is close to $\pm1$ remains to be studied. The Lagrange interpolation theorem with nodes $x_j$ takes the form

$$P(x) = \sum_1^m P(x_j)\frac{T_m(x)}{T_m'(x_j)(x-x_j)} = \frac{1}{m}\sum_1^m P(x_j)\frac{T_m(x)(-1)^{j-1}\sin t_j}{x-x_j},$$

where $t_j = \frac{(2j-1)}{m}\frac{\pi}{2}$, for it is easy to prove that $T_m'(x_j) = (-1)^{j-1}m/\sin t_j$. Say for definiteness that $x_1 < x \leqslant 1$. Then since $|P(x_j)\sin t_j| \leqslant M/m$ and all the differences $x - x_j$ are positive, we obtain

$$|P(x)| \leqslant \frac{M}{m^2}\sum_1^m \left|\frac{T_m(x)}{x-x_j}\right| = \frac{M}{m^2}\left|\sum_1^m \frac{T_m(x)}{x-x_j}\right|.$$

The last sum is just the derivative of $T_m$ (think of the logarithmic derivative of $T_m$) and we note that $T_m'(\cos t) = m\sin mt/\sin t$ is bounded by $m^2$ in absolute value. Hence $|P(x)| \leqslant M$ also for these values of $x$.

**Lemma 10.3.** *Let $S$ be an odd trigonometric polynomial of order $m$. Then*

$$\|S/\sin\| \leqslant m\|S\|.$$

*Proof.* Using the well-known formula

$$\frac{\sin(k+1)t}{\sin t} = \cos kt + \frac{\sin kt}{\sin t}\cos t$$

we can prove by induction that $\sin kt/\sin t$ is a polynomial in $\cos t$ of order $k-1$. Hence $S(t) = P(\cos t)\sin t$ for some algebraic polynomial $P$ of degree $m-1$. We now apply Lemma 10.2 to $P$ to conclude.

*Proof of Theorem 10.1.* Consider the function $f(s,t) = \frac{1}{2}(T(s+t) - T(s-t))$. For fixed $s$, the trigonometric polynomial $t \mapsto f(s,t)$ is odd and of order at most $m$. Clearly $|f(s,t)| \leqslant \|T\|$, so by Lemma 10.3, $|f(s,t)/\sin t| \leqslant m\|T\|$. Hence also $T'(s) = \lim_{t\to0}(f(s,t)/\sin t)(\sin t/t)$ is bounded by $m\|T\|$ in modulus.

**Theorem 10.4** (Markov's[16] inequality). *Every polynomial $P$ of degree $m$ satisfies $\|P'\|_\infty \leqslant m^2\|P\|_\infty$, where the norms are supremum norms on $[-1,1]$.*

*Proof.* The function $T(t) = P(\cos t)$ is a trigonometric polynomial of order $m$, so Bernstein's inequality implies $|T'(t)| = |P'(\cos t)\sin t| \leqslant m\|P\|$, which can be written $|P'(x)\sqrt{1-x^2}| \leqslant m\|P\|$ for $x \in [-1,1]$. We now apply Lemma 10.2 to $P'$ to conclude: $\|P'\| \leqslant m^2\|P\|$.

For any $\alpha$ with $0 < \alpha \leqslant 1$, let us define a class $\mathrm{Lip}_\alpha(\mathbf{R})$ of functions, consisting of those whose modulus of continuity satisfies $\omega(\delta) = O(\delta^\alpha)$, $\delta \to 0$. The elements of $\mathrm{Lip}_1(\mathbf{R})$ are called *Lipschitz functions.*

**Theorem 10.5.** *Let $f \in C_{2\pi}(\mathbf{R})$ and fix $0 < \alpha < 1$. Then $E_m(f) = O(m^{-\alpha})$ as $m \to \infty$ if and only if $f \in \mathrm{Lip}_\alpha(\mathbf{R})$.*

*Proof.* That functions in $\mathrm{Lip}_\alpha$ can be approximated as indicated is Jackson's first theorem. Assume conversely that we know that $E_m(f) = O(m^{-\alpha})$. Thus there exists trigonometric polynomials $T_m$ of order at most $m$ such that $\|f - T_m\| \leqslant Cm^{-\alpha}$, $m \geqslant 1$, for some constant $C$ independent of $m$. The convergence of the sequence $(T_m)$ to $f$ is in general too slow for a telescoping series to be useful. We shall take a subsequence $(T_{2^j})$, which converges faster. By the triangle inequality we have

$$|f(x+t) - f(x)| \leqslant |f(x+t) - T_{2^k}(x+t)| + |T_{2^k}(x+t) - T_{2^k}(x)| + |T_{2^k}(x) - f(x)|.$$

Here the first and third terms on the right do not exceed $C2^{-\alpha k}$, and the second term can be estimated by $|t|\|T'_{2^k}\|$. To estimate this derivative we shall use the telescoping series

$$T'_{2^k} = T'_1 + \sum_1^k (T'_{2^j} - T'_{2^{j-1}}),$$

together with the estimate

$$(10.2)\quad \|T_{2^j} - T_{2^{j-1}}\| \leqslant \|T_{2^j} - f\| + \|f - T_{2^{j-1}}\| \leqslant C(2^{-\alpha j} + 2^{-\alpha(j-1)}) = C_1 2^{-\alpha j}.$$

The last inequality implies in view of Bernstein's inequality that

$$\|T'_{2^j} - T'_{2^{j-1}}\| \leqslant C_1 2^j 2^{-\alpha j},$$

so that

$$\|T'_{2^k}\| \leqslant C_2 + \sum_1^k C_1 2^{j(1-\alpha)} \leqslant C_2 + C_3 2^{k(1-\alpha)},$$

where $C_2$ is a bound for $T'_1$. Putting things together we have

$$|f(x+t) - f(x)| \leqslant 2C2^{-\alpha k} + |t|\left(C_2 + C_3 2^{k(1-\alpha)}\right),$$

where $k$ is still to be chosen. Given $t$ with $|t| \leqslant 1$ we pick $k$ such that $2^k \leqslant 1/|t| < 2^{k+1}$ and see that $|f(x+t) - f(x)| \leqslant C_4|t|^\alpha$ for a new constant $C_4$, $|t| \leqslant 1$. For larger $|t|$ this is true at the expense of a larger constant. This completes the proof.

**Exercise 10.6.** Prove that if $f$ is a periodic function with $E_m(f) = O(1/m)$, then its modulus of continuity satisfies $\omega(\delta) \leqslant C\delta \log \delta$, $0 < \delta < \frac{1}{2}$, for some constant $C$.

It is to be noted that the proof of Theorem 10.5 breaks down for $\alpha = 1$. Indeed the result is not true in that case. Instead, we have the following theorem.

---

[16] Andrej Andreevič Markov, 1856–1922.

**Theorem 10.7** (Zygmund's[17] Theorem). *Let $f \in C_{2\pi}(\mathbf{R})$ be given. Then $E_m(f) = O(1/m)$, $m \to \infty$, if and only if $\omega_2(\delta) = O(\delta)$, $\delta \to 0$.*

**Exercise 10.8.** Construct a function such that $\omega_2(\delta) = O(\delta)$, $\delta \to 0$, but which is not Lipschitz.

*Proof of Theorem 10.7.* Theorem 7.12 proves one direction. To prove the other direction, assume that $E_m(f) = O(1/m)$, $m \to \infty$, and write $f = \lim T_{2^j}$, where $T_{2^j}$ is a trigonometric polynomial of order at most $2^j$ such that $\|f - T_{2^j}\| \leqslant C2^{-j}$. Thus we can use a telescoping series

$$f = \lim_{k\to\infty} T_{2^k} = T_1 + \sum_1^\infty (T_{2^j} - T_{2^{j-1}}) = \sum_0^\infty V_j, \tag{10.3}$$

where $V_0 = T_1$ and $V_j = T_{2^j} - T_{2^{j-1}}$, $j \geqslant 1$. The second difference of $f$ can then be estimated by

$$\left|\Delta_t^2 f\right| \leqslant \sum_0^{m-1} \left|\Delta_t^2 V_j\right| + 4\sum_m^\infty \|V_j\|,$$

where we still can choose $m$. From (10.2) we see that $\|V_j\| \leqslant C_1 2^{-j}$, $j \geqslant 1$. The second difference is easy to estimate by the second derivative, so we can conclude, using Bernstein's inequality, that for $|t| \leqslant \delta$,

$$\left|\Delta_t^2 f\right| \leqslant \sum_0^{m-1} \delta^2 \|V_j''\| + C_2 2^{-m} \leqslant \sum_0^{m-1} \delta^2 2^{2j} \|V_j\| + C_2 2^{-m} \leqslant C_1 \delta^2 2^m + C_2 2^{-m}$$

when $|t| \leqslant \delta$. We now choose $m$ so that $2^m \leqslant \delta^{-1} < 2^{m+1}$ and get the desired result.

**Theorem 10.9** (Bernstein). *Let $f \in C_{2\pi}(\mathbf{R})$ and fix $k \in \mathbf{N}$ and $\alpha \in [0,1[$. Then*

$$E_m(f) \leqslant Cm^{-k-\alpha}, \qquad m \in \mathbf{N},$$

*if and only if $f \in C^k(\mathbf{R})$ and $f^{(k)} \in \mathrm{Lip}_\alpha(\mathbf{R})$.*

*Proof.* We shall prove only one direction: we assume that $f$ admits the approximation $E_m(f) \leqslant Cm^{-k-\alpha}$ and shall prove that $f^{(k)}$ exists and is in $\mathrm{Lip}_\alpha(\mathbf{R})$. We have a representation $f = \sum_0^\infty V_j$ as in (10.3) where now

$$\|V_j\| \leqslant \|T_{2^j} - f\| + \|f - T_{2^{j-1}}\| \leqslant C(2^j)^{-k-\alpha} + C(2^{j-1})^{-k-\alpha} = C_1 2^{-j(k+\alpha)}.$$

From Bernstein's inequality we deduce that

$$\|V_j^{(k)}\| \leqslant (2^j)^k \|V_j\| \leqslant C_1 2^{jk} 2^{-jk-j\alpha} = C_1 2^{-j\alpha}.$$

Hence $\sum V_j^{(k)}$ converges uniformly to some continuous function $g$, and since termwise integration is allowed we see that $f^{(k)}$ exists and is equal to $g$. Now

$$E_{2^j}(f^{(k)}) \leqslant \left\| f^{(k)} - \textstyle\sum_0^j V_i^{(k)} \right\|,$$

[17] Antoni Zygmund, 1900–1992.

for $\sum_0^j V_i^{(k)}$ is a trigonometric polynomial of order at most $2^j$. However,

$$\left\| f^{(k)} - \sum_0^j V_i^{(k)} \right\| = \left\| \sum_{j+1}^\infty V_i^{(k)} \right\| \leqslant C_2 2^{-j\alpha},$$

which shows that $E_{2^j}(f^{(k)}) \leqslant C_2 2^{-j\alpha}$. For any $m \in \mathbf{N}$ we now choose $j$ such that $2^j \leqslant m < 2^{j+1}$, which yields the estimate

$$E_m(f^{(k)}) \leqslant E_{2^j}(f^{(k)}) \leqslant C_2 2^{-j\alpha} \leqslant 2^\alpha C_2 m^{-\alpha}.$$

Now Theorem 10.5 implies that $f^{(k)} \in \mathrm{Lip}_\alpha(\mathbf{R})$.

## References

Baker, George A., Jr., and Gammel, John L. (Eds.)
1970 *The Padé Approximant in Theoretical Physics.* New York and London: Academic Press. xii + 378 pp.

Baker, George A., Jr., and Graves-Morris, Peter
1981 *Padé Approximants. Part I: Basic Theory*; *Part II: Extensions and Applications.* Encyclopedia of Mathematics and its Applications, Volumes 13 and 14. Reading, MA: Addison-Wesley. xx + 325 resp. xviii + 215 pp.

Bernstein, Serge
1912 Démonstration du théorème de Weierstrass, fondée sur le calcul des probabilités. *Commun. Soc. Math. Kharkow* (2) **13** (1912–13), 1–2.
1913 Sur la meilleure approximation de $|x|$ par des polynômes de degrés donnés. *Acta Math.* **37**, 1–57.

Borel, É.
1905 *Leçons sur les fonctions de variables réelles.* Paris.

Chebyshev, P. L.
1854 Teorija mehanizmov, izvestnyh pod nazvaniem parallelogrammov. [Collected works 1948, vol. 2, 23–51.]

Cheney, E. W.
1966 *Introduction to Approximation Theory.* McGraw-Hill. xii + 259 pp.

Dzjadyk, V. K.
1977 *Vvedenie v teoriju ravnomernogo približenija funkcij polinomami.* Moscow: Nauka.

Haar, Alfred
1918 Die Minkowskische Geometrie und die Annäherung an stetige Funktionen. *Math. Ann.* **78**, 293–311.

Jackson, D.
1912 On approximation by trigonometric sums and polynomials. *Trans. Amer. Math. Soc.* **13**, 491–515.

Kolmogorov, A. N.
1948 Zamečanie po povodu mnogočlenov P. L. Čebyševa, naimenee uklonjajuščihsja ot zadannoj funkcii. *Uspehi mat. nauk* **3**, 216–221.

Kornejčuk, N. P.

1962 The exact constant in D. Jackson's theorem on best uniform approximation of continuous periodic functions. [Russian.] *Doklady Akad. Nauk* **145**, 514–515.

Korovkin, P. P.

1959 *Linejnye operatory i teorija približenij.* Moscow: Gosudarstvennoe izdatel′stvo fiziko-matematičeskoj literatury. English translation: *Linear operators and Approximation Theory.* Delhi: Hindustan Publishing Company, 1960.

Landau, E. G. H.

1908 Über die Approximation einer stetigen Funktion. *Rendiconti Circ. mat. Palermo* **25**, 337–345.

Lorentz, G. G.

1953 *Bernstein polynomials.* Toronto: University of Toronto Press.

Lozinskij, S. M.

1948 On a class of linear operators. [Russian.] *Doklady Akad. Nauk* **61**, 193–196.

Mairhuber, J.

1956 On Haar's theorem concerning Chebysheff approximation problems having unique solutions. *Proc. Amer. Math. Soc.* **7**, 609–615.

Newman, D.

1964 Rational approximation to $|x|$. *Michigan Math. J.* **11**, 11–14.

Rivlin, T. J.

1990 *Chebyshev Polynomials.* 2nd ed. New York: Wiley.

Rudälv, Catarina

1998 *Multivariate rational approximation with prescribed poles.* Umeå University, Department of Mathematics, Doctoral Thesis No. 16.

Stahl, Herbert

1992 Best uniform rational approximation of $|x|$ on $[-1,1]$. [Russian.] *Mat. Sb.* **183**, 85–118.

Stone, M. H.

1937 Applications of the theory of Boolean rings to general topology. *Trans. Amer. Math. Soc.* **41**, 375–481.

1948 The generalized Weierstrass approximation theorem. *Math. Mag.* **21**, 167–184.

Weierstraß, Karl

1885 Über die analytische Darstellung sogenannter willkürlicher Funktionen einer rellen Veränderlichen. *Sitzungsberichte der Akad. zu Berlin*, 633–639, 789–805.

Author's address: Uppsala University, Department of Mathematics,
P. O. Box 480, SE-751 06 Uppsala, Sweden.

Telephone: +46 18 4713216 (office); +46 18 300708 (home) Fax: +46 18 4713201

Electronic mail: kiselman@math.uu.se

URL: http://www.math.uu.se/~kiselman