Duality Methods in Čebyšev Approximation

Pták, Vlastimil
1997

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Duality Methods in Čebyšev Approximation

Vlastimil Pták

Technical report No. 708

April 1997

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+4202) 6884244 fax: (+4202) 8585789
e-mail: uivt@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Duality Methods in Čebyšev Approximation

Vlastimil Pták

Technical report No. 708
April 1997

Abstract

An expository text explaining how duality methods may be used to obtain the classical theorems on Čebyšev approximation as consequences of a simple result on representation of functionals on finite-dimensional spaces of continuous functions. The representation lemma is based, in its turn, on the classical theorem of Carathéodory on convex hulls. In this manner the classical theorems of Haar, Čebyšev and de la Vallée-Poussin are reduced to the basic facts on convexity.

Keywords

The proof is based on a representation theorem for linear functionals on finite - dimensional spaces of continuous functions and this, in its turn, uses the following basic facts from convexity theory.

(1.1) *If E is a (not necessarily closed) subspace f a Banach space X and $x_0 \in X$ then*

$$\text{dist}\,(x_0, E) = \max\left\{\langle x, f\rangle; f \in E^0, |f| \leq 1\right\}$$

(1.2) *If A is a compact subset of an n-dimensional Banach space E then the absolutely convex bull of A is compact and consists of all vectors of the form $\sum_{i=1}^{n} \lambda_i a_i$ with $a_i \in A$ and $\sum_{i=1}^{n} |\lambda_i| \leq 1$.*

The second fact is closely related to the classical theorem of Carathéodory which says that every element of the convex hull of a set A in an n-dimensional vector space is a convex combination of $n+1$ points a_i in A.

Let T be a compact Hausdorff space and let $C(T)$ be the Banach space of all continuous realvalued functions on T equipped with the norm

$$|x| = \max\{|x(t)|;\ t \in T\}.$$

The dual space of $C(T)$ will be denoted by $C(T)^d$, the value of the functional $f \in C(T)^d$ at $x \in C(T)$ will be denoted by $\langle x, f\rangle$. Given $t \in T$, the mapping $x \longmapsto x(t)$ clearly is linear and bounded and, consequently, may be considered as an element of $C(T)^d$. We shall denote this functional by $\varphi(t)$; thus $\langle x, \varphi(t)\rangle = x(t)$ for $x \in C(T)$ and $t \in T$. Whenever there is no danger of misunderstanding, we shall identify t and $\varphi(t)$ and write simply $\langle x, t\rangle = x(t)$. It will be convenient to form linear combinations of functionals of this type: in this identification $f = \sum \lambda_j t_j$ means

$$\langle x, f\rangle = \sum \lambda_j x(t_j)$$

for every $x \in C(T)$.

The problem considered by A.Haar may be stated (in modern terminology) as follows. Given an n-dimensional subspace $E \subset C(T)$ a simple compactness argument shows that, for every $x \in C(T)$, there exists at least one $e_0 \in E$ such that $|x - e_0| = \inf\{|x - e|; e \in E\}$. This element e_0, the best approximation of x by elements of E, will not be unique in general. In the case where T is a compact interval of the real axis and E is the subspace of all polynomial of degree $\leq n-1$ (restrictions of polynomials to T) we do have uniqueness so we can speak about *the* polynomial of best approximation or the Čebyšev polynomial. In the case of a general n-dimensional subspace E we have the problem of A.Haar: to characterize those n-dimensional subspaces $E \subset C(T)$ such that for each $x \in C(T)$, the best approximation by elements of E is unique. The result of Haar may be stated in the form of the following equivalence.

(1.3) Theorem *Let E be an n-dimensional subspace of $C(T)$. Then the best approximation by elements of E is unique for every $x \in C(T)$ if and only if no nonzero element $e \in E$ can have more than $n-1$ distinct zeros in T.*

This explains, in particular, why best approximations by polynomials are unique. Indeed, a polynomial of degree $\leq n-1$ with more than $n-1$ distinct zeros is identically zero.

We intend to show now that the substance of the approximation theorems is contained in a representation theorem for linear functionals on n-dimensional subspaces of $C(T)$. It is a known fact that bounded linear functionals on $C(T)$ may be represented by certain measures on T. The following basic result shows that linear functionals on n-dimensional subspaces of $C(T)$ have a representation by measures concentrated in n points of T.

Before stating the theorem it will be convenient to introduce some notation. Consider a Banach space E and let L be a subspace of E. Given $f \in E^d$, the norm of its restriction to L will be denoted by $|f|_L$; clearly $|f|_L \leq |f|$ and the inequality will be strict except in the case when the supremum

$$\sup\langle x, f\rangle; \quad x \in E, \quad |x| \leq 1$$

may already be reached taking elements $x \in L$ only.

(1.4) Representation theorem *Let L be a p-dimensional subspace of $C(T)$. Let f be a linear functional on L. Then there exist p points $t_i \in T$ and numbers $\lambda_1, \ldots, \lambda_p$ such that $f = \sum_{i=1}^{p} \lambda_i t_i$ and $\sum |\lambda_i| = |f|_L$.*

Proof. With every point $t \in T$ we may associate a linear functional $\psi(t)$ on L defined by the relation $\langle x, \psi(t)\rangle = x(t)$ for every $x \in L$. In this manner $\psi(t)$ is the restriction to L of the functional $\varphi(t) \in C(T)^d$ so that its norm $|\psi(t)|_L \leq |\varphi(t)| = 1$ for every $t \in T$. It follows that $\psi(T)$ is a subset of the closed unit bull B of L^d. Since ψ is clearly a continuous mapping of T into B the set $\psi(T)$ is compact. Let us show now that B coincides with the absolutely convex hull of $\psi(T)$. Indeed, suppose there exists a $b \in B$ that does not belong to the absolutely convex hull of $\psi(T)$. Since linear functionals on L^d may be identified with elements of L we conclude that there exists an $x \in L$ such that

$$\sup\langle x, \psi(T)\rangle < \langle x, b\rangle.$$

This is a contradiction since

$$\sup\langle x, \psi(T)\rangle = |x| \quad \text{and} \quad \langle x, b\rangle \leq |x||b| \leq |x|.$$

According to the theorem of Carathéodory (its strengthened version for absolutely convex hulls) every element g of B may be represented in the form $g = \sum_{i=1}^{p} \lambda_i \psi(t_i)$ with $\sum_{i=1}^{p} |\lambda_i| \leq 1$.

Now consider an arbitrary nonzero $f \in L^d$; the functional $g = |f|_L^{-1} f$ is an an element of B and may thus be represented in the form

$$g = \sum_{i=1}^{p} \alpha_i \psi(t_i) \quad \text{with} \quad \sum |\alpha_i| \leq 1.$$

Since $1 = |g|_L \leq \sum |\alpha_i|$ we have $\sum |\alpha_i| = 1$. Setting $\lambda_i = |f|_L \alpha_i$ we obtain the desired representation of f. □

Now we are ready to prove the theorem of Haar. It will be convenient to restate it in its negative form.

(2.1) Theorem *Let E be an n-dimensional subspace of $C(T)$. The the following two conditions are equivalent.*
1° there exists an $x \in C(T)$ for which the best approximation by elements of E is not unique.
2° there exists a nonzero $e \in E$ such that $e(t) = 0$ for at least n distinct points $t \in T$.
Proof. Elementary linear algebra shows that condition 2° is equivalent to
3° *there exists a nonzero linear combination $\sum_{i=1}^{n} \lambda_i t_i$ which vanishes on E.*

Now assume 1°. Then there exist elements $x_0 \in C(T)$, $e_0 \in E$ and a nonzero $e \in E$ such that, for every ε sufficiently small in modulus,

$$|x_0 - (e_0 + \varepsilon e)| = \text{dist}\,(x_0, E).$$

Denote by L the linear span of E and x_0 so that the dimension of L is $n + 1$. There exists a linear functional f of norm one on L such that f vanishes on E and

$$\langle x_0 - e_0, f\rangle = |x_0 - e_0|$$

By the representation theorem f may be expressed in the form $f = \sum_{i=1}^{n+1} \lambda_i t_i$ with $\sum |\lambda_i| = 1$. If one of the λ_i is zero condition 3° is satisfied and 3° implies 2°. We may thus assume that the t_i are distinct and all the $n+1$ coefficients λ_i are different from zero. Writing w_0 for $x_0 - e_0$ we obtain $|w_0| = \langle w_0, f\rangle = \sum_{i=1}^{n+1} \lambda_i w_0(t_i)$. Since $\sum |\lambda_i| \leq 1$ and all $\lambda_i \neq 0$, it follows that $|w_0(t_i)| = |w_0|$ for all i. Furthermore, $|w_0 + \varepsilon e| = |w_0|$ for all ε sufficiently small in modulus. Fixing i, we have

$$|w_0(t_i) + \varepsilon e(t_i)| \leq |w_0 + \varepsilon e| = |w_0| = |w_0(t_i)|$$

for ε small in modulus; this is not possible unless $e(t_i) = 0$. Thus $e(t_i) = 0$ for all the $n + 1$ points t_i and this gives 2°.

On the other hand, suppose there exists a nonzero $e_0 \in E$, $|e_0| \leq 1$ that vanishes at n distinct points $t_i \in T$. It follows that there exists a nonzero linear combination $f = \sum_{i=1}^{n} \lambda_i t_i$ such that $\langle E, f\rangle = 0$. We may clearly suppose that $\sum_{i=1}^{n} |\lambda_i| = 1$ so that $|f| \leq 1$. Now choose an arbitrary $a \in C(T)$ such that $|a| \leq 1$ and $a(t_i) = \text{sign}\ \lambda_i$ whenever $\lambda_i \neq 0$ and define the function x_0 by the relation $x_0(t) = a(t)(1 - |e_0(t)|)$. It follows that $|x_0(t)| + |e_0(t)| \leq 1$ for every $t \in T$ whence $|x_0 - e_0| \leq 1$; at the same time, $|x_0| = 1$. If we show that $|x_0 - e| \geq 1$ for all $e \in E$ it will follow that 0 and e_0 are two different best approximations of x_0. Now, for every $e \in E$,

$$|x_0 - e| \geq |f||x_0 - e| \geq \langle x_0 - e, f\rangle = \langle x_0, f\rangle = 1$$

and this completes the proof. □

Now suppose E is an n-dimensional subspace of $C(T)$ which satisfies Haar's condition so that, for every $x \in C(T)$ there exists exactly one $e_0 \in E$ such that

$$|x - e_0| = \inf\{|x - e|; e \in E\}.$$

Thus e_0 is the best approximation of x by elements of E. The representation theorem may be used to obtain additional information about the set of those points $t \in T$ where the modulus of the function $x(t) - e_0(t)$ assumes its maximal value $|x - e_0|$. Indeed, we have the following result.

(2.2) *Let E be an n-dimensional subspace of $C(T)$. Suppose that E satisfies Haar's condition. If $x \in C(T)$ lies outside E and if e_0 is the best approximation of x by elements of E then the equation*

$$|x(t) - e_0(t)| = |x - e_0|$$

has at least $n + 1$ distinct solution.

Proof. Since x does not belong to E the linear span L of E and x is a linear subspace of dimension $n + 1$. There exists a linear functional f of norm one on L such that f vanishes on E and

$$\langle x - e_0, f\rangle = |x - e_0|$$

Now f may be expressed in the form $f = \sum_{i=1}^{n+1} \lambda_i t_i$ with $t_i \in T$ and $\sum |\lambda_i| = 1$. Since E satisfies Haar's condition the t_i must be all distinct and the λ_i must all be different from zero. Since

$$|x - e_0| = \langle x - e_0, f\rangle = \sum_{i=1}^{n+1} \lambda_i(x(t_i) - e_0(t_i))$$

it follows that $|x(t_i) - e_0(t_i)| = |x - e_0|$ for every i. □

In the case where T is a compact interval of the real axis it is possible to prove a more precise statement including information about the sign of $x(t)-e_0(t)$ at the points t where the norm of $x - e_0$ is attained; this will enable us to give a characterization of the element of best approximation. Let us show now how the classical theorems of Čebyšev and de la Valée-Poussin may be obtained as immediate consequences of the representation theorem (1.4). We begin by proving a simple lemma.

(2.3) *Let T be a compact interval of the real axis and let E be an n-dimensional subspace of $C(T)$ which satisfies Haar's condition. If $t_1 < t_2 < \ldots < t_{n+1}$, are given points of T there exists exactly one (up to a scalar factor) nonzero linear combination $f = \sum_{i=1}^{n+1} \lambda_i t_i$ vanishing on E. All numbers λ_i are different from zero and they alternate in sign.*

Proof. Since E is an n-dimensional space, the $n+1$ linear functionals t_j are linearly dependent on E so that there exists a linear combination $f = \sum_{i=1}^{n+1} \lambda_i t_i$ with $\langle E, f\rangle = 0$. Our assumption about E implies that all λ_i are different from zero.

Now suppose that two consecutive coefficients λ are of the same sign, positive say. We have thus, for some i, the inequalities $\lambda_i > 0$, $\lambda_{i+1} > 0$. Since E has dimension n there exists an $e \in E$ such that $e(t_i) = 1$ and $e(t_j) = 0$ for all j different from i and $i+1$, if any, i.e. if $n \geq 2$. Let us show now that $e(t_{i+1}) > 0$. Indeed, if $e(t_{i+1}) = 0$, the element e has at least n distinct zeros on T. The inequality $e(t_{i+1}) < 0$ would imply, together with $e(t_i) > 0$, the existence of at between t_i and t_{i+1} for which $e(t) = 0$. In both cases we obtain a contradiction with Haar's condition for E. Thus $e(t_{i+1}) > 0$ and

$$\langle e, f\rangle = \lambda_i e(t_i) + \lambda_{i+1} e(t_{i+1}) > 0$$

a contradiction. The proof is complete. □

The following lemma is due to de la Vallée-Poussin - we obtain it as an immediate consequence of the preceding lemma.

(2.4) *Let T be a compact interval of the real axis and let E be an n-dimensional subspace of $C(T)$ fulfilling Haar's condition. Let $x_0 \in C(T)$ and $e_0 \in E$ be given. Suppose $t_1 < t_2 < \ldots < t_{n+1}$ are given points in T. Suppose that the numbers $x_0(t_i) - e_0(t_i)$ are all different from zero and alternate in sign. Then*

$$\min\{|x_0 - e|; e \in E\} \geq \min |x_0(t_j) - e_0(t_j)|$$

Proof. According to the preceding lemma there exist positive numbers $\lambda_f, \ldots, \lambda_{n+1}$ with $\sum_{i=1}^{n+1} \lambda_i = 1$ such that $f = \sum_{i=1}^{n+1} (-1)^i \lambda_i t_i$ vanishes on E. Set $\varepsilon = \operatorname{sign}(x_0(t_1) - e_0(t_1))$. For every $e \in E$ we have then

$$\begin{aligned} |x_0 - e| &\geq \langle x_0 - e, \varepsilon f\rangle = \langle x_0 - e_0, \varepsilon f\rangle = \\ &= \varepsilon \sum \lambda_i (-1)^i (x_0(t_i) - e_0(t_i)) = \sum \lambda_i |x_0(t_i) - e_0(t_i)| \geq \\ &\geq \min |x_0(t_j) - e_0(t_j)| \end{aligned}$$

□

Now we are ready to prove the theorem of Čebyšev.

(2.5) Theorem *Let T be a compact interval of the real axis and let E be an n-dimensional subspace of $C(T)$ fulfilling Haar's condition. Let $x \in C(T)$ and $e \in E$ be given. Then the following two conditions are equivalent.*

1° *The element e is the best approximation of x by elements of E*

2° *There exist $n+1$ points $t_1 < t_2 < \ldots < t_{n+1}$ in T such that $x(t_i) - e(t_i)$ are equal to $|x - e|$ in modulus and alternate in sign.*

Proof. Assume 1°. Then there exists a functional $f = \sum_{i=1}^{n+1} \lambda_i t_i$ with $\sum_{i=1}^{n+1} |\lambda_i| = 1$ which vanishes on E and such that $|x - e| = \langle x - e, f\rangle$. It follows that $x(t_i) - e(t_i) = |x - e| \text{ sign } \lambda_i$ whenever $\lambda_i \neq 0$. If the t_i are arranged in increasing order it follows from lemma (2.3) that the corresponding λ are all different from zero and alternate in sign. This proves 2°. The implication 2° → 1° follows from the preceding result. □

(3.1) Example. *Let T be a set of cardinality $n > 1$ and consider the $(n-1)$ dimensional subspace E of $C(T)$ consisting of those $x \in C(T)$ for which $\sum_{t \in T} x(t) = 0$. Clearly E satisfies the Haar condition. Let x_0 be defined by $x_0(t) = 1$ for all $t \in T$. Then 0 is the best approximation of x_0 by elements of E. In particular, the element of best approximation e_0 satisfies $x_0(t) - e_0(t) = 1$ for all $t \in T$.*

Proof. It suffices to prove the following implication: if $\sum x_i = 0$ then $\max |x_i - 1| \geq 1$. Suppose, on the contrary, that this maximum is less than 1. Then $\sum |x_i - 1| < n$ so that

$$0 > \sum |x_i - 1| - n = \sum (|x_i - 1| - 1) \geq \sum \left((x_i - 1)^2 - 1\right) = \sum x_i^2 \geq 0$$

a contradiction.

Bibliography

[1] V. Pták: A remark on approximation of continuous functions, Czech. Math. J. 83 (1958), 251-256

[2] V. Pták: On the absolutely convex hull of a convex set in a finite-dimensional vector space (in Czech), Čas. pěst. mat. 83 (1958), 343 - 347