ELSEVIER

Journal of Computational and Applied Mathematics 121 (2000) 1–36

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.nl/locate/cam

Approximation in normed linear spaces

G.A. Watson

Department of Mathematics, University of Dundee, Dundee DD1 4HN, Scotland, UK

Received 10 October 1999; received in revised form 10 January 2000

Abstract

A historical account is given of the development of methods for solving approximation problems set in normed linear spaces. Approximation of both real functions and real data is considered, with particular reference to L_p (or l_p) and Chebyshev norms. As well as coverage of methods for the usual linear problems, an account is given of the development of methods for approximation by functions which are nonlinear in the free parameters, and special attention is paid to some particular nonlinear approximating families.

1. Introduction

The purpose of this paper is to give a historical account of the development of numerical methods for a range of problems in best approximation, that is problems which involve the minimization of a norm. A treatment is given of approximation of both real functions and data. For the approximation of functions, the emphasis is on the use of the Chebyshev norm, while for data approximation, we consider a wider range of criteria, including the other l_p norms, $1 \leqslant p < \infty$. As well as the usual linear problems, a general account is given of nonlinear best approximation, and we also consider some special cases. Only a passing mention is made of least-squares problems, as that is considered elsewhere. The focus is also entirely on the approximation of real quantities, and so best approximation of complex quantities is not covered. A partial justification of this is that dealing with problems in generality as complex ones would introduce additional complication not entirely justified by the additional algorithmic initiatives.

Since we are concerned here with historical development, technical details are not included for their own sake. The intention is, where appropriate, to be descriptive, rather to give a technically rigorous and detailed account of methods. However, it seemed necessary at times for the sake of comprehensibility, and in order to fully appreciate algorithmic developments, to include a reasonable amount of technical detail.

E-mail address: gawatson@maths.dundee.ac.uk (G.A. Watson).

0377-0427/00/$ - see front matter
PII: S0377-0427(00)00333-2

Obviously a major factor in the development of methods has been the advent of powerful computing facilities, as this has opened up opportunities to tackle a wide range of practical problems. Whereas at one time, the main consideration may have been elegance and simplicity, with attention perhaps focussed on a set of problems satisfying "classical" assumptions, those considerations now usually have to take second place to the treatment of problems which are seen to be of practical importance, for which algorithms have to be robust and efficient.

The paper is effectively divided into two parts, the first (Section 2) being concerned with approximation by linear families, and the second (Section 3) being concerned with approximation by nonlinear families. These sections themselves further subdivide into two parts, where we consider separately approximation of data and of functions, and these are dealt with in that order within the two sections, with a further breakdown in what seems to be a reasonably natural way to take account of important special cases.

For the approximation of functions, we are primarily concerned with univariate functions on an interval $[a,b]$, because that is where most effort has been concentrated. However, some relevant comments are made on the extent to which multivariate functions may also be treated, with a few references made to this.

2. Linear approximation

The approximation of a given function defined on an interval by a linear combination of given functions is the most fundamental problem in approximation theory. The functions involved are usually continuous, and this can be thought of as a continuous infinite dimensional approximation problem. If the functions are replaced by vectors in $\mathbb{R}^m$, then we have a class of finite dimensional or discrete problems, many of which have their origins in data fitting. That solutions to linear best approximation problems always exist is a result which goes back at least to Riesz in 1918 [174]. We will consider the finite dimensional problem first, and begin by making some general remarks, before looking at special cases.

2.1. Linear approximation in $\mathbb{R}^m$

Let $A \in \mathbb{R}^{m\times n}$ where $m \geqslant n$, and let $\boldsymbol{b} \in \mathbb{R}^m$. Then the statement of a linear best approximation problem in $\mathbb{R}^m$ can be given as

$$\text{find } \boldsymbol{x} \in \mathbb{R}^n \text{ to minimize } ||\boldsymbol{r}||, \tag{1}$$

where

$$\boldsymbol{r} = A\boldsymbol{x} - \boldsymbol{b},$$

and $||.||$ is a given norm on $\mathbb{R}^m$. The dependence of $\boldsymbol{r}$ on $\boldsymbol{x}$ will generally be suppressed, unless confusion is possible.

This particular problem has attracted enormous interest. It will be assumed throughout that $\text{rank}(A) = n$, and there is no $\boldsymbol{x}$ such that $\boldsymbol{r} = 0$. These are not essential, neither in theory nor in practice; however, they are conditions that are normally satisfied in practice, and their assumption considerably simplifies the presentation. If the norm is a differentiable function of $\boldsymbol{x}$, then we can easily characterize a minimum by zero derivative conditions: these are necessary, and, exploiting

convexity, also sufficient. The best known example is when the norm is the least-squares norm, when zero derivative conditions just give the usual normal equations

$$A^{\mathrm{T}}A\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}.$$

The method of least squares is considered in detail elsewhere. But in a data fitting context, other l_p norms, particularly those for values of p satisfying $1 \leqslant p < 2$ are also important. The reason for this is that it is common for the usual conditions justifying the use of the l_2 norm not to hold, for example there may be wild points or gross errors in the data, and these other norms give reduced weight to these wild points. This is considered in Sections 2.2 and 2.3. Of great interest also has been the use of the Chebyshev norm; this is perhaps of less value in a data fitting context, but problems arise for example in continuous function approximation when the region of approximation is discretized. The problem is rich in structure and the theory is a beautiful one; we consider this case in Section 2.4.

We will restrict attention here to the problem (1), although there are many modifications of that problem which are relevant in a data fitting context. Most modifications have only been given serious treatment comparatively recently, and so they are of lesser interest from a historical point of view.

2.2. Linear l_1 approximation in $\mathbb{R}^m$

Consider now the problem (1) with the l_1 norm

$$||\boldsymbol{r}||_1 = \sum_{i=1}^{m} |r_i|. \tag{2}$$

This problem has a long history: its statement goes back well into the mid eighteenth century, and predates the introduction of least squares. Certainly, it was used in work of Laplace in 1786, in solving the overdetermined system of linear equations determining planetary orbits [110]. The first systematic methods for solving this problem seem due to Edgeworth [61]; in 1887 he gave a method based on tabulation, and in 1888 a method for the case when $n = 2$ which was essentially graphical and conceptual, but based on calculating descent directions. In 1930, Rhodes [167], motivated by the problem of fitting a parabola to data, tried Edgeworth's later method but found it "cumbrous". He gave a method where each iteration was calculated by solving 2 interpolation conditions for 2 of the parameters, and minimizing with respect to the remaining parameter. A proof that this kind of approach can give a solution was established by Singleton in 1940 [182]. A detailed historical account is given by Farebrother in a 1987 paper [63], covering the period 1793 to 1930.[1]

The first modern systematic study of this problem appears to be by Motzkin and Walsh [131,132] in the late 1950s, and characterization results are given in the 1964 book by Rice [172]. A convenient form of these may be deduced from these results or as a simple consequence of applying to this special case known results in abstract approximation theory: we will not attempt to go down that historical route, since it is something of a diversion from the main theme. However, it is the case

[1] The 1999 book by Farebrother [64] is also relevant.

that a vector $\boldsymbol{x} \in \mathbb{R}^n$ solves the l_1 problem if and only if there exists a vector $\boldsymbol{v} \in \mathbb{R}^m$ satisfying

$$A^{\mathrm{T}}\boldsymbol{v} = 0,$$

where $||\boldsymbol{v}||_\infty \leqslant 1$, and $v_i = \operatorname{sign}(r_i)$ whenever $r_i \neq 0$. The first simple (direct) proof of this was probably given by Watson [199] in 1980. A formal treatment of the important result that when A has rank n, a solution will be such that n components of r_i are zero, was given by Motzkin and Walsh [131] in 1955. In the context of the l_1 problem, any point characterized in this way can be defined to be a *vertex*. The interpolation result (in special cases) appears to have been known to Gauss, and to have been used in early methods: for example, the methods of Rhodes and Singleton are essentially vertex to vertex descent methods.

The results of Motzkin and Walsh were arrived at by direct consideration of the problem. However, its relationship with a linear programming problem was recognized around the same time,[2] and linear programming theory provides a parallel route to the same properties. Around 1947, Dantzig did his pioneering work on the simplex method of linear programming, and over the next few years, duality theory was developed, largely by von Neumann, Gale, Kuhn and Tucker. The significance of these developments for numerical methods for the l_1 (and the l_∞) problem cannot be overemphasized.

The first representation of the l_1 problem as a tractable linear programming problem seems due to Charnes et al. [35] in 1955. The key observation is that if extra variables $\boldsymbol{u}$ and $\boldsymbol{v} \in \mathbb{R}^m$ are introduced, then the problem can be posed as

$$\text{minimize } \sum_{i=1}^{m} (u_i + v_i) \text{ subject to} \tag{3}$$

$$[I : -I : A] \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{x} \end{bmatrix} = \boldsymbol{b}$$

$$\boldsymbol{u} \geqslant 0, \quad \boldsymbol{v} \geqslant 0.$$

Since in the simplex method, no columns of I and $-I$ can simultaneously be basic, then

$$u_i v_i = 0, \quad i = 1, \ldots, m.$$

It follows that $u_i + v_i = |u_i - v_i|$ for all i and the equivalence of the simplex method applied to this problem with the minimization of (2) can readily be established.

Another version of the primal can be stated:

$$\text{minimize } \boldsymbol{e}^{\mathrm{T}}\boldsymbol{s} \text{ subject to}$$

$$-\boldsymbol{s} \leqslant A\boldsymbol{x} - \boldsymbol{b} \leqslant \boldsymbol{s}.$$

This goes back at least to the 1964 Russian edition of the book by Zuhovitskii and Avdeyeva [211]. However, this form of the problem does not seem to have attracted as much attention as (3). The zero residuals will result in a form of degeneracy.

[2] Farebrother [63] in his 1987 paper interprets the work of Edgeworth in this context, and states that "..it must be conceded that Edgeworth had developed a fully operational, if somewhat complex, linear programming procedure for the L_1 estimation problem in 1888".

Fisher [66] in 1961 gave some publicity to (3) for the benefit of the statistical community, and this form was also used by Barrodale and Young [13] in 1966, who provided an Algol implementation and numerical results. The fact that the components of $\boldsymbol{x}$ may be non-negative is not a major problem in this context: for example, they can each be replaced by the difference of two non-negative variables. It was also noted that no first phase simplex calculation is required because an initial basic feasible solution can readily be obtained: if $b_i < 0$ then $\boldsymbol{e}_i$ can be present in the initial basis, if $b_i > 0$ then $-\boldsymbol{e}_i$ can be, with either used if $b_i = 0$.

The linear programming connection is sometimes wrongly credited to Wagner [192] in 1959, who posed the problem as a bounded variable or interval programming problem. In fact the form of the problem considered by Wagner [192] can be interpreted as the dual of (3). This can be written as

$$\text{maximize } \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v} \text{ subject to} \tag{4}$$

$$A^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}$$

$$-\boldsymbol{e} \leqslant \boldsymbol{v} \leqslant \boldsymbol{e},$$

where $\boldsymbol{e}$ is a vector with every component equal to 1. Attention was re-focussed on (4) by Robers and Ben-Israel [175] in 1969, and Robers and Robers [176] in 1973, who argued the advantages of that approach, which included computational efficiency: the problem with the primal appeared to be the large number of extra variables required. However, an improved version of the primal linear programming method was given by Davies [53] in 1967 and Barrodale and Roberts [10] in 1970, where a special pivot column selection rule was employed, and in 1973, both Spyropoulos et al. [183] and Barrodale and Roberts [11] gave efficient implementations of the simplex method applied to the primal which fully exploited the structure. The Barrodale and Roberts method achieved efficiency by taking multiple pivot steps, exploiting the fact that descent can continue beyond the usual point when feasibility is lost, because feasibility can readily be recovered by swapping certain variables into and out of the basis. Further efficiency was achieved by imposing certain restrictions on the choice of variables to enter and leave the basis. A Fortran programme and numerical results were provided, together with favourable comparisons with some other primal and dual methods [12].

In 1975, Abdelmalik [2] developed a special purpose method for the dual, using the dual simplex method, and his method seemed comparable with that of Barrodale and Roberts [10]. This turned out not really to be surprising, because, as pointed out by Armstrong and Godfrey [6] in 1979, the application of the dual simplex method to the dual is equivalent to applying the primal simplex method to the primal. So apart from implementation aspects, the methods were the same.

A basic feasible solution to (3) in which all columns of A are present in the basis can readily be shown to correspond to a vertex as defined above. Therefore, once the columns of A are present in the basis, the simplex method is a vertex to vertex descent method. There are many other variants of these linear programming methods, but away from a linear programming context, *direct* descent methods were being considered. For given $\boldsymbol{x}$, let

$$Z = \{i : r_i = 0\}.$$

Then since for full rank problems the solution occurs at a point $\boldsymbol{x}$ with Z containing n indices (a vertex), we want to systematically descend to such a point. Perhaps the first modern direct descent methods were given by Usow [189] in 1967, and Claerbout and Muir [43] in 1973. A natural way

to implement descent methods is by first finding a vertex, and then descending through a sequence of vertices. Thus there are two types of step depending on whether at the current point, Z contains (a) fewer than n indices (b) exactly n indices. (The possibility that Z contains more than n indices corresponds to a degenerate situation, and although there are ways round it, will for our purposes be ignored.) Then in case (a) movement as far as possible is made in the direction $\boldsymbol{d}$ in such a way that the number of indices in Z at the new point is increased, and in case (b) movement as far as possible is made in the direction $\boldsymbol{d}$ in such a way that the number of indices in Z is maintained. Effective methods of this type, therefore, have this strategy in common, and are distinguished by the way the descent direction is calculated. There are mainly two approaches, (i) *reduced gradient methods*, where the "active constraints" are used to express certain variables in terms of others, the objective function is expressed in terms of the latter group, and its gradient is obtained in terms of those, and (ii) *projected gradient methods*, where the gradient is obtained by projecting the gradient of the objective function onto the orthogonal complement of the span of the active constraints.

Bartels et al. [18] in 1978 gave a projected gradient method, and reduced gradient methods were given by Osborne [147,148] in 1985 and 1987. Both projected and reduced gradient methods were analyzed in detail by Osborne [147] in 1985, and he pointed out that although reduced gradient methods seem more suitable for implementation using a tableau format, with updating, in fact such organization is available for implementing both methods. On relationships with linear programming methods, he showed that there is an exact equivalence between the possible options available in implementing the simplex method and those available in the direct application of the reduced gradient method. Thus these algorithms are equivalent: only the implementational details are different. The usual simplex step corresponds to a particular option in the reduced gradient method, based on an unnormalized steepest edge test for determining the variable to leave the basis. A different way of choosing this variable (a normalized steepest edge test, which is scale invariant) was used by Bloomfield and Stieger [26] in 1983, and their evidence showed that this can lead to improvement.

Nearly all the good methods considered to the end of the 1980s were vertex to vertex methods, which exploit the polyhedral nature of the function to be minimized, and (in the absence of degeneracy) they are finite. There has been recent interest in interior point methods for linear programming problems, stimulated by the results of Karmarker [102] in 1984. In conjunction with a formal connection with classical barrier methods for constrained optimization problems, this has resulted in renewed interest in linear programming, and there has of course been an impact on special cases such as the l_1 problem.

The use of interior point methods for l_1 problems goes back at least as far as work of Meketon [127] in 1987, and methods have been given since then by Ruzinsky and Olsen [178] in 1989, Zhang in 1993 [209] and Duarte and Vanderbei [56] in 1994. Portnoy and Koenker [157] in 1997 make a case for the superiority of interior point methods over simplex-based methods for large problems. Based on comparisons of l_1 problems having n up to 16 and m from 1000 to 200 000, they conclude that there is "a compelling general case for the superiority of interior point methods over traditional simplex methods for large linear programming problems". Their algorithm of choice for the l_1 problem is based on a primal–dual log barrier method due to Mehrotra [123] in 1992, and includes a statistical preprocessing approach which estimates whether a residual is zero or not. The opposition is represented by a variant of the Barrodale and Roberts method.

Meantime, two other types of smoothing method were being developed for the l_1 problem.[3] The first of these is typified by an algorithm of Coleman and Li [46] in 1992, which is based on affine scaling: while not strictly speaking an interior point method, it is nevertheless in the spirit of such methods. Here, an attempt is made to satisfy the characterization conditions by an iterative descent method which has the following characteristics: (a) it generates a sequence of points which are such that Z is empty, so that derivatives exist, (b) it is globally convergent, (c) it ultimately takes damped Newton steps (damped to satisfy (a)), but with sufficiently accurate approximations to the full Newton step to permit quadratic convergence (under nondegeneracy conditions). Careful implementation of the method can avoid difficulties with near-zero components of $\boldsymbol{r}$ and the approach seems promising for large problems as it is insensitive to problem size. Some comparisons show that it is superior to Meketon's interior point method for problems with n up to 200, m up to 1000.

A second approach to smoothing the l_1 problem was developed by Madsen and Nielsen [116] in 1993. It is based on the use of the Huber M-estimator, defined by

$$\psi_\gamma \equiv \psi_\gamma(\boldsymbol{r}) = \sum_{i=1}^{m} \rho(r_i), \tag{5}$$

where

$$\rho(t) = \begin{cases} t^2/2, & |t| \leqslant \gamma, \\ \gamma(|t| - \gamma/2), & |t| > \gamma, \end{cases} \tag{6}$$

and γ is a scale factor or tuning constant. The function (5) is convex and once continuously differentiable, but has discontinuous second derivatives at points where $|r_i| = \gamma$. The mathematical structure of the Huber M-estimator seems first to have been considered in detail by Clark [44] in 1985. Clearly if γ is chosen large enough, then ψ_γ is just the least-squares function; in addition if γ tends to zero, then limit points of the set of solutions may be shown to minimize the l_1 norm. It is the latter property which concerns us here.

It has been suggested by Madsen and Nielsen [116] in 1993 and also by Li and Swetits [113] in 1998 that the preferred method for solving the l_1 problem is via a sequence of Huber problems for a sequence of scale values $\gamma \to 0$. This algorithmic development has lead to increased interest in the relationship between the Huber M-estimator and the l_1 problem; for example there is recent work of Madsen et al. [117] in 1994, and Li and Swetits [113] in 1998. The method of Madsen and Nielsen generates Huber solutions for a sequence of values of γ, tending to zero. The solutions are obtained by solving least-square problems, exploiting structure so that new solutions can be obtained using updating often in $O(n^2)$ operations. A key feature is that it is not necessary to let γ reach zero; once a sufficiently small value is identified, then the l_1 solution may be obtained by solving an $n \times n$ linear system. Madsen and Neilsen give some comparisons (for randomly generated problems, and with m mostly set to $2n$ for m up to 1620) with the method of Barrodale and Roberts [10] and claim superiority.

An important issue as far as the implementation of simplex type methods is the efficiency of the line search. The Barrodale and Roberts [10] method incorporates the equivalent of a comparison sort, and this leaves room for considerable improvement. Bloomfield and Stieger [26] considered

[3] The observation that a best approximation can always be computed as the limit of a sequence of l_p approximations as $p \to 1$ is due to Fischer [65] in 1983 (an algorithm based on this was in fact given by Abdelmalik [1] in 1971), although this is not a very practical approach.

this aspect in their 1983 book, and suggested using a fast median method. An alternative based on the use of the secant algorithm was considered (in a related context) by George and Osborne [71] in 1990, and again by Osborne [147] in 1985. Numerical experiments were reported by Osborne and Watson [154] in 1996, where the secant-based method was seen to be as good as fast median methods on randomly generated problems, and to perform considerably better on problems with systematic data. Comparisons of other types of method with simplex methods really need to take this into account before definitive conclusions can be drawn.

2.3. Linear l_p approximation in $\mathbb{R}^m$, $1 < p < \infty$, $p \neq 2$

For given $\boldsymbol{x} \in \mathbb{R}^n$, let $D_{|r|}$ be defined by

$$D_{|\boldsymbol{r}|} = \operatorname{diag}\{|r_1|, \ldots, |r_m|\}.$$

Then $\boldsymbol{x}$ minimizes

$$||\boldsymbol{r}||_p^p = \sum_{i=1}^{m} |r_i|^p$$

with $1 < p < \infty$ if and only if derivatives with respect to $\boldsymbol{x}$ are zero, that is if

$$A^{\mathrm{T}} D_{|r|}^{p-1} \theta = 0, \tag{7}$$

where $\theta_i = \operatorname{sign}(r_i)$, $i = 1, \ldots, m$. This is a nonlinear system of equations for $\boldsymbol{x}$.

This criterion (for p even) was mentioned by Gauss as a generalization of his least-squares criterion. Apart from this special case, the more general l_p problem only seems to have attracted relatively recent computational attention. The range $1 < p < 2$ is of particular interest computationally because there is potentially reduced smoothness: problems with $p \geqslant 2$ are twice differentiable, but problems with $1 < p < 2$ may be only once differentiable. If $p \geqslant 2$ or if $1 < p < 2$ and no component of $\boldsymbol{r}$ is zero then twice differentiability is guaranteed and so (7) can be written as

$$A^{\mathrm{T}} D \boldsymbol{r} = 0, \tag{8}$$

where

$$D = D_{|\boldsymbol{r}|}^{p-2},$$

and this is a particularly convenient form with which to work. It represents a generalized system of normal equations, effectively a least-squares problem apart from the "weighting" matrix D. Fixing $\boldsymbol{x}$ to an approximate value in D and solving this weighted system for a new approximation gives an example of the technique known as *iteratively reweighted least squares* or IRLS, which seems to have been introduced by Beaton and Tukey [20] in 1974. Since good software for (weighted) least-squares problems was then available, this seemed an attractive idea, additionally so since there are some apparently good theoretical properties: this simple iteration process will converge locally if p is close to 2, and if zero components of $\boldsymbol{r}$ are avoided, it is globally convergent (from any initial approximation) for $1 < p < 2$. The last result seems first to have been given by Dutter [60] in 1975. However, convergence can be slow, particularly as p nears 1 (it is linear with convergence constant $|p-2|$, as shown by Wolfe [206] in 1979), and there are potential numerical difficulties for reasons which will be clear from the previous section. The matrix D (which may not exist) can

be replaced by approximations (even by the unit matrix), and this gives rise to variants of the IRLS technique, but again convergence can be very slow.

Most recent algorithms for solving (8) are based on Newton's method, and many variants were proposed in the 1970s. It is interesting that the Newton step is just $1/(p-1)$ times the IRLS step (as measured by the difference between successive approximations), as pointed out by Watson [196] in 1977, and this gave an explanation of some success obtained by Merle and Späth [128] in 1974 in using a damped IRLS procedure with step length $(p-1)$. Thus apart from differences due to the line search, IRLS and Newton's method with line searches *are essentially the same method.* It is easily seen that the region of convergence of Newton's method is proportional to $|(p-1)/(p-2)|$, so good line search procedures are needed even with the basic method, certainly far from $p=2$. However, for $p>2$, Newton's method with line search is usually perfectly satisfactory.

Since from a practical point of views the interesting cases are those when $1<p<2$, different strategies have been proposed for getting round the difficulties arising from zero (or near zero) components of $\boldsymbol{r}$. These included the substitution of small nonzero values, solving a slightly perturbed problem, or identifying and so removing these components from the set. However, not just zero components but *nearly zero* components are potentially troublesome. There is some evidence, however, that these phenomena are not *by themselves* a major problem, but only if they are accompanied by p being close to 1. The main difficulty appears to be due the fact that as p approaches 1, we are coming closer to a discontinuous problem, effectively to a constrained problem. It seems necessary to recognize this in a satisfactory algorithm, and consider some of the elements of the l_1 problem in devising an approach which will deal in a satisfactory with small values of p. This is the philosophy in a recent method due to Li [114] in 1993, which is essentially equivalent to the method for the l_1 problem of Coleman and Li [46] referred to in the previous section. Numerical results show that the new method is clearly superior to IRLS (with the same line search) for values of p close to 1, with the gap between the two methods widening as p approaches 1. There is little difference for values of $p \geqslant 1.5$ or so. As with the l_1 case, the number of iterations appears to be independent of the problem size.

2.4. Linear Chebyshev approximation in $\mathbb{R}^m$

The use of the criterion now known as the Chebyshev norm

$$||\boldsymbol{r}||_\infty = \max_i |r_i|, \tag{9}$$

seems to go back to Laplace in 1786, who gave a solution procedure for $n=2$. Cauchy in 1814 and Fourier in 1824 gave descent methods. A detailed historical account is given by Farebrother [63] in his 1987 paper, covering the period 1793 to 1824. The function space analogue was studied first by Chebyshev[4] from the 1850s, arising from an analysis of a steam engine linking, and both continuous and discrete problems now carry his name.

For any $\boldsymbol{x} \in \mathbb{R}^n$, let

$$\bar{I}(\boldsymbol{x}) = \{i\colon |r_i(\boldsymbol{x})| = ||\boldsymbol{r}||_\infty\}.$$

[4] The number of variants in the western literature which have been used for Chebyshev is legendary, but most people now seemed to have settled on this one. Butzer and Jongmans [33] in 1999 gave a detailed account of Chebyshev's life and work.

Then $\boldsymbol{x}$ is a solution if and only if there exists a subset $I \subset \bar{I}$ containing at most $n+1$ indices, and a nontrivial vector $\lambda \in \mathbb{R}^m$ such that

$$\lambda_i = 0, \quad i \notin I,$$

$$A^{\mathrm{T}}\lambda = 0,$$

$$\lambda_i \operatorname{sign}(r_i) \geqslant 0, \quad i \in I.$$

This is an example of a "zero in the convex hull" type of characterization result, and is the discrete analogue of Kirchburger's 1903 result for the continuous problem [106]. A simple consequence is that for the full rank problem, there always exists a solution with $\bar{I}$ containing $n+1$ indices. Such a point can be thought of as a *vertex*.

An early method for the minimization of (9) was the Polya algorithm [156], which solves a sequence of l_p problems with $p \to \infty$: the assumption here is that the l_p problems are relatively easy to solve, being differentiable for large finite p. This method was given (in fact for continuous functions) in 1913, and convergence is readily established if the Chebyshev solution is unique. A proof of convergence to a particular Chebyshev approximation called the strict Chebyshev approximation (in the event of nonuniqueness of the Chebyshev solution) was given by Descloux [54] in 1963. Fletcher et al. [69] in 1974 used an extrapolation technique to accelerate convergence of the Polya algorithm, and in the same year Boggs [27] used a technique based on deriving a differential equation describing the l_p solution as a function of p. An algorithm due to Lawson [111] in 1961 was based on the solution of a sequence of weighted least-squares problems, but like the Polya algorithm, it can be very slowly convergent. Indeed none of these methods has been regarded as giving a particularly practical approach.

A fundamental assumption which was identified as important at an early stage was the *Haar condition*, that every $n \times n$ submatrix of A is nonsingular. This is sufficient for uniqueness of $\boldsymbol{x}$ minimizing (9) (and also necessary in the case when $m = n+1$). This "classical" assumption played a major role in the minimization of (9) until the 1960s. It goes back to Haar [79] in 1918.

Before proceeding, it is helpful to point out an important property which is satisfied at a minimum of (9). The result, due to de la Vallée Poussin [190] in 1911, tells us that if J runs through all subsets of $n+1$ indices from $\{1,\ldots,m\}$, then

$$\min_{\boldsymbol{x}} \max_i |r_i| = \max_J \left\{ \min_{\boldsymbol{x}} \max_{i \in J} |r_i| \right\}. \tag{10}$$

Therefore, if we can identify a set J where the maximum on the right-hand side is attained, solving a Chebyshev problem *on that subset* (and this is relatively easy) will give a solution to the original problem. For any J such that the corresponding problem matrix is full rank, the solution on J will occur at a vertex. Therefore, if A has full rank, so that the problem has a solution at a vertex, then it is sufficient to investigate all the vertices in a systematic way.

An exchange algorithm for finding an extremal subset or optimal vertex was given by Stiefel [184] in 1959. It assumed that A satisfied the Haar condition, and worked with a sequence $J_1, J_2, \ldots$ of subsets of $n+1$ components of $\boldsymbol{r}$. The key aspect of the method was that J_{k+1} differed from J_k by one index, and a rule was given for exchanging one of the indices in J_k for another index outside it to give J_{k+1} in such a way that

$$h_{k+1} > h_k,$$

where

$$r_i(\boldsymbol{x}) = \theta_i h_k, \quad i \in J_k, \tag{11}$$

with $|\theta_i| = 1,\ i \in J_k$. Thus, we have an example of a vertex-to-vertex *ascent method*. Because there are only a finite number of selections of $n+1$ indices from m the method must terminate, when from (10) it follows that a solution has been obtained. If the Haar condition is not satisfied, then strict inequality may hold for successive h_k values and the theory of the method is compromised.

Because the function of $\boldsymbol{x}$ given by (9) is piecewise linear, the Chebyshev problem may be posed as a linear programming problem, and as in the l_1 case, properties of the problem are again available through this route. Let $h = \max_i |r_i|$. Then the problem may be stated

minimize h subject to

$$-h \leqslant r_i \leqslant h, \quad i = 1, \ldots, m.$$

In terms of the variables h and $\boldsymbol{x}$, this may be restated

$$\text{minimize } z = \boldsymbol{e}_{n+1}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{x} \\ h \end{bmatrix} \text{ subject to}$$

$$\begin{bmatrix} A & \boldsymbol{e} \\ -A & \boldsymbol{e} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ h \end{bmatrix} \geqslant \begin{bmatrix} \boldsymbol{b} \\ -\boldsymbol{b} \end{bmatrix}.$$

One of the first to consider the linear programming formulation of the problem was Zuhovickii [211,212] in a series of papers originating in the Russian literature in the early 1950s. The above form is not particularly suitable for the application of standard techniques such as the simplex method because $2m$ slack variables are required, the basis matrices will be $2m \times 2m$, and although h is nonnegative, this is not true in general of the components of $\boldsymbol{x}$.

All of these difficulties are overcome by turning to the dual problem, which is

$$\text{maximize } z = [\boldsymbol{b}^{\mathrm{T}},\ -\boldsymbol{b}^{\mathrm{T}}]\boldsymbol{w} \text{ subject to}$$

$$[A^{\mathrm{T}} \quad -A^{\mathrm{T}}]\boldsymbol{w} = 0,$$

$$[\boldsymbol{e}^{\mathrm{T}} \quad \boldsymbol{e}^{\mathrm{T}}]\boldsymbol{w} \leqslant 1,$$

$$\boldsymbol{w} \geqslant 0.$$

Only one slack variable is required (to make the inequality an equality), the basis matrices are only $(n+1) \times (n+1)$, and all the variables are nonnegative. The advantage in using the dual seems to have been first pointed out by Kelley in [105] 1958 in an application to curve fitting. Standard linear programming theory tells us that if a variable is dual basic, then the corresponding primal constraint holds with equality. Thus a basic feasible solution corresponds precisely to a solution to a set of equations having the form (11). It would appear therefore that there is a precise equivalence between a step of the simplex method applied to the dual, and a step of the Stiefel exchange method. This result was known to Stiefel [185] in 1960, who gave an indication of it by considering a small problem and using a geometric argument. He also (unnecessarily) eliminated the unconstrained variables from the primal before proceeding to the dual.

The precise equivalence was first worked out in detail by Osborne and Watson [149] in 1967, although Bittner [24] in 1961 examined how linear programming could be used to relax the Haar condition. In order for the usual simplex method to be applied to the dual, the Haar condition is not required, only the nonsingularity of successive basis matrices: for this it is only necessary for A to have full rank n. The point here is that the simplex method does not permit nonsingular basis matrices. Note however that if the Haar condition does not hold, degeneracy may cause cycling in the simplex algorithm, although this can always be resolved.

A programme implementing the standard simplex method for the problem considered here seems first to have been given by Barrodale and Young [13] in 1966, who gave an Algol programme along with numerical results, and Bartels and Golub [19] gave a version in 1968 which used a numerically stable factorization procedure. In 1975, Barrodale and Phillips [8] used the special structure present in the dual formulation to greatly reduce the number of iterations required: conditions were imposed on variables entering and leaving the basis, and the fact exploited that some variables could easily be exchanged for others. The usual simplex rules were modified to permit ascent through a number of vertices, beyond the one which would usually be reached in a simplex step, by exploiting the fact that feasibility could easily be regained by such exchanges. Modifications of this basic technique to allow more than one index to be exchanged at each step were given by Hopper and Powell [90] in 1977 and by Armstrong and Kung [7] in 1979.

The Stiefel exchange method and variants which solve the dual formulation of the problem are examples of ascent methods, whose justification is based on (10). However, it is possible to solve the problem by a *descent* process. The primal linear programming problem is an example of a descent method, and although its direct solution is not recommended for the reasons already given, it is nevertheless possible to implement satisfactory descent methods.

As for the l_1 problem, good direct descent methods might be expected to follow the common strategy of having (in the absence of degeneracy) basically two types of step depending on whether the current point $\boldsymbol{x}$ is such that $\bar{I}$ contains (a) fewer than $n+1$ indices, or (b) exactly $n+1$ indices. In a manner precisely analogous to that considered for the l_1 problem, a strategy can be developed which ultimately gives a vertex-to-vertex descent process. Methods of reduced gradient type were given by Cheney and Goldstein [37] in 1959 and Cline [45] in 1976. A projected gradient method was given by Bartels et al. [16] in 1978. It appeared to be the case that such methods did not seriously compete with ascent methods. However, improvements in descent methods were considered by Bartels et al. [17] in 1989: they argued that the good performance of the Barrodale and Phillips method was due to the way the method chose a good starting point. By modifying the way in which a starting point is obtained for their descent method, they enhanced its performance and made a case for its superiority for data fitting problems.

All the approaches considered so far are essentially vertex-to-vertex methods. They exploit the polyhedral nature of the function to be minimized, and are of course (in the absence of degeneracy) finite. The recent interest in interior point methods for linear programming problems has, as in the l_1 case, extended to the special case of Chebyshev problems. Ruzinsky and Olsen [178] in 1989, Zhang [209] in 1993 and Duarte and Vanderbei [56] in 1994 all proposed interior point methods. An affine scaling algorithm analogous to that for the l_1 problem was given by Coleman and Li [47] in 1992. This is a descent method which involves a sequence of least-squares problems to define descent directions. It provides a smooth transition from guaranteed descent steps far from a solution, to steps close to a solution which are sufficiently accurate

approximations to the Newton step to permit quadratic convergence under suitable nondegeneracy assumptions.

In contrast to the l_1 situation, detailed comparisons of other methods with simplex type methods for large problems do not yet seem to be available. It should in any event not be assumed that conclusions can be drawn from the l_1 case, because large Chebyshev problems normally arise from discretizations of continuous Chebyshev approximation problems on intervals or multidimensional regions, and the data are highly systematic. Indeed, the solution is then normally part of a method for the continuous problem, or exploits the connection: we will defer further consideration of this until the following Section.

2.5. Linear Chebyshev approximation in $C[a,b]$

Let $C[a,b]$ denote the set of continuous functions defined on the real interval $[a,b]$, and let $f(x)$, $\phi_1(x),\ldots,\phi_n(x)$ be in $C[a,b]$. Then the usual Chebyshev approximation problem in $C[a,b]$ can be expressed as

$$\text{find } \boldsymbol{a} \in \mathbb{R}^n \text{ to minimize } ||f-\phi||_\infty, \tag{12}$$

where $\phi = \sum_{i=1}^n a_i\phi_i(x)$, and

$$||f||_\infty = \max_{a\leqslant x\leqslant b} |f(x)|.$$

This class of problems was systematically investigated by Chebyshev from the 1850s, although Chebyshev credits Poncelet with originating the problem. The "classical" case occurs when the set of functions forms a Chebyshev set (or is a Haar subspace) on $[a,b]$, that is any nontrivial linear combination has at most $(n-1)$ zeros; the model problem here is approximation by polynomials of degree $n-1$. The problem (12), with the interval $[a,b]$ replaced by m points in $[a,b]$, reduces to a problem of the form considered in the previous section. Indeed it is readily seen that the matrix A in this case satisfies the Haar condition if and only if the set of functions $\phi_1(x),\ldots,\phi_n(x)$ forms a Chebyshev set on $[a,b]$. Continuing this theme for a moment, arbitrarily good solutions to (12) can be obtained by choosing finer and finer discretizations; the main convergence results here are due to Motzkin and Walsh [132] in 1956. Although this observation by itself does not give practical algorithms, the use of a sequence of discretizations, where successive point sets are carefully chosen, is the key to the success of many good algorithms.

A general characterization result was obtained by Kirchberger [106] in 1903. Let

$$\bar{E} = \{x \in [a,b], |r(x,\boldsymbol{a})| = ||r(.,\boldsymbol{a})||_\infty\}.$$

Then $\boldsymbol{a}$ is a solution if and only if there exists $E \subset \bar{E}$ containing $t \leqslant n+1$ points $x_1,\ldots,x_t$ and a nontrivial vector $\lambda \in \mathbb{R}^t$ such that

$$\sum_{i=1}^t \lambda_i\phi_j(x_i) = 0, \quad j = 1,\ldots,n,$$

$$\lambda_i \operatorname{sign}(r_i) \geqslant 0, \quad i = 1,\ldots,t.$$

Borel [28] in 1905 established the well-known alternation result for approximation by degree $(n-1)$ polynomials, that $\boldsymbol{a}$ is a solution if and only if there are $n+1$ points in $[a,b]$ where the

norm is attained with alternating sign as we move from left to right through the points; we can state this concisely in the form

$$\mathscr{A}(f - p_{n-1})_{[a,b]} \geqslant n + 1,$$

where p_n denotes the best degree n polynomial approximation. Uniqueness of solutions under these conditions is also due to Borel [28] in 1905. That this result extends to approximation by functions forming a Chebyshev set was shown by Young [208] in 1907, who also established uniqueness in this case. Haar [79] in 1918 showed that the solution is unique for all possible functions $f(x)$ if and only if $\phi_1(x), \ldots, \phi_n(x)$ forms a Chebyshev set on $[a, b]$.

Polya [156] in 1913 gave his algorithm for this problem, where a sequence of continuous L_p problems is solved with $p \to \infty$. A counterexample to a general convergence result for non-Chebyshev set problems was given by Descloux [54] in 1963. As in the discrete case this is anyway not a particularly effective approach.

Two important algorithms for solving the Chebyshev problem were given by Remes [165,166] in the 1930s. The method traditionally known as the "Second Algorithm" applies to Chebyshev set problems, exploiting the alternation property. It solves a sequence of discrete problems in $\mathbb{R}^{n+1}$ defined by sets of $n+1$ points in $[a, b]$: each of these is just the solution of a system of $n+1$ equations for $n + 1$ unknowns, using the fact the solutions have an alternation property. By exchanging the current set of $n + 1$ points for $n + 1$ local maxima of the modulus of the error function, subject to some simple rules, an ascent process is obtained. Under mild conditions this converges to the (unique) Chebyshev approximation, and at a second-order rate: the result, due to Veidinger [191] in 1960, is based on showing that the method is asymptotically Newton's method for solving the characterization conditions. A comparatively modern implementation of the method was given by Golub and Smith [73] in 1971. Note that if only one point is exchanged at each iteration (bringing in a point where the norm of the error is attained), then an equivalence can be drawn between a step of the method and a step of the Stiefel exchange method. An analysis of the one-point exchange method is given by Powell [158] in his 1980 book, where it is shown that this method also converges at a second-order rate.

The "First Algorithm of Remes" applies to general problems. Again it corresponds to the solution of a sequence of discrete problems, but of increasing size. Starting with a solution on $m_1 \geqslant (n + 1)$ discrete points in $[a, b]$, a point where the error function attains the norm is added, and a new solution obtained on $m_1 + 1$ points. If the matrix A of the initial problem has rank n, then successive matrices also have rank n and so linear programming techniques, for example, can be used, and implemented efficiently using postoptimality theory. This is an "implicit" exchange method, since every solution corresponds to a vertex defined on the current set of points. In fact since much of the work in implementing such a method lies in finding a global maximum of the error function, and this would normally involve calculating all the local maxima, it is sensible to add in all such local maxima: the method is then an implicit multiple exchange method. For Chebyshev set problems this is equivalent to the second algorithm of Remes. Modifications of the first algorithm of Remes to allow multiple exchanges have been considered by Carasso and Laurent [34] in 1978, and Blatt [25] in 1984, based on constructing "chains of references".

Unfortunately there are examples where this kind of approach performs badly, when the solution to the continuous problem does not occur at a vertex, that is it attains the norm in fewer than $n + 1$ points: such problems were called singular problems by Osborne and Watson [151] in 1969. Note

that this phenomenon is specific to problems on a continuum, and has no analogue in the (full rank) discrete case. Therefore, because each discrete problem has a solution at a vertex, the limiting situation in this case is obtained by some of these points coalescing, slowing down convergence and giving ill-conditioned simplex bases.

For multivariate problems (where x is a vector in $\mathbb{R}^s$, $s > 1$), singularity is very common. A partial explanation for this is that Chebyshev sets of more than one function do not exist in continuums of dimension higher than one: this was first pointed out by Mairhuber [119] in 1956. Nevertheless, a method of this type can be developed for multivariate problems, as demonstrated by Watson [194] in 1975.

Therefore, there are two main difficulties with such methods: (a) the calculation of the local and global maxima of the error function, (b) the problem of singularity. It is perhaps only recently that close attention has been paid to efficient calculation in (a), for example by Reemtsen [163] in 1991, and Price and Coope [159] in 1996: it is usually assumed that all local maxima can be calculated to sufficient accuracy, and so the relevant algorithms are always implementable. But attempts to avoid (a) have been made, for example by Dunham [58] in 1981, Hettich [85] in 1986 and Reemtsen [162] in 1990. The main idea is to only require maxima of the error at each step on a grid, where the discretization error tends to zero as the method progresses. In particular, Reemtsen proved the convergence of a modified version of the first algorithm of Remes, in which the maximum of the kth error function was computed on a grid, with the grid density tending to zero. The method of Hettich is also based on successive grid refinement (and using a numerically stable simplex algorithm) and applies to one- and two-dimensional problems; solutions have been successfully obtained for problems with n up to 37.

An alternative approach which tries to avoid both (a) and (b) is through the use of *two-phase methods*. The first phase involves the solution of a single discretization of the problem, on a sufficiently dense set to enable identification of the number of points (with signs) where the norm is attained and good approximations of these. In the second phase, the characterization conditions, together with zero derivative conditions at points identified as extrema in (a, b), can then be solved (for example by Newton's method). This main idea for an approach of this type (in a more general context) is due to Gustafson [77] in 1970. Its application to Chebyshev approximation problems was considered by a number of people in the mid-1970s, among them Gustafson, Hettich, Andreassen and Watson [5,78,84]. The approach can be successful, but while the difficulty (a) above is essentially removed, (b) can still emerge in the first phase, and there is also the (new) difficulty of having to decide what level of discretization to use, or when to enter the second phase, and also when the information provided at that point is completely reliable. It may be necessary to permit re-entry to phase 1 with a more stringent exit criterion.

The second phase can be considered in two ways, depending on whether or not the local maxima are considered as differentiable functions of the unknown parameters, and whether or not this is exploited. If it is, then the zero derivative conditions can be used to eliminate these maxima in terms of the other unknowns, and there is a consequent reduction in the size of the linear system to be solved for the Newton step.

Of course the second phase applies equally to nonlinear problems, so we will return to some of these ideas in Section 3.4. Indeed, continuous Chebyshev approximation problems (both linear and nonlinear) are special cases of semi-infinite programming problems, that is problems with a finite number of variables and an infinite number of constraints, and many algorithms which have

been developed for the more general problem class may be adapted for the Chebyshev approximation problem. Semi-infinite programming is an active research area – the recent survey paper of Reemtsen and Görner [164] in 1998 has 233 references, 96 of them dated 1990 or later. Algorithmic development has encompassed methods based on the ideas considered above, but also other approaches, for example the use of interior point methods. These are of comparatively recent origin, their usefulness (certainly as far as continuous Chebyshev approximation is concerned) does not appear to have been established, and we will not consider them further here.

2.6. Chebyshev approximation by splines with fixed knots

Approximation by splines is considered in some detail elsewhere, so we will not go into the history of the origins of this class of function. The main focus of approximation by splines has been on interpolation; however, Chebyshev approximation by splines has also attracted a lot of attention. Because we are concerned at present with linear problems, we assume in the present section that the knots are fixed a priori, and we will consider approximation from the space of spline functions defined as follows. Let integers m and k be given, and let $a = x_0 < x_1 < \cdots < x_{k+1} = b$. Then

$$S_m = \{s \in C^{m-1}[a,b] : s(x) \in \Pi_m \text{ on } [x_i, x_{i+1}],\ i = 0, \ldots, k\},$$

where Π_m denotes the space of polynomials of degree m, is the space of polynomial splines of degree m with k fixed knots. S_m is a linear space with dimension $m + k + 1$. The first results on Chebyshev approximation by splines seem to be due to Johnson [96] in 1960.

The theory of approximation by Chebyshev sets does not apply to approximation from S_m. However, S_m is an example of a family of functions forming a *weak Chebyshev set*: any linear combination of such a set of n functions has at most $(n-1)$ changes of sign. For such sets Jones and Karlowitz [100] showed in 1970 that there exists at least one best Chebyshev approximation ϕ to any continuous function f which has the classical alternation property

$$\mathscr{A}(f - \phi)_{[a,b]} \geqslant n + 1,$$

(although there may be others which do not).

The theory of Chebyshev approximation by splines with fixed knots is fully developed, and a characterization of best approximation goes back to the Ph.D. dissertation of Schumaker in 1965, and his publications over the next few years, e.g. [180]. Results were also given by Rice [173] in 1967. What is required is the existence of an interval $[x_p, x_{p+q}] \subset [a,b]$, with $q \geqslant 1$ such that there are at least $q + m + 1$ alternating extrema on $[x_p, x_{p+q}]$, or in the notation previously introduced

$$\mathscr{A}(f - s)_{[x_p, x_{p+q}]} \geqslant q + m + 1,$$

where $s \in S_m$. In addition to characterization of solutions, there has been interest in conditions for uniqueness (and strong uniqueness) of best approximations. In general of course, best approximations are not unique. However, the uniqueness (and strong uniqueness) of best spline approximations is characterized by the fact that all knot intervals contain sufficiently many alternating extrema as shown by Nürnberger and Singer [143] in 1982.

The solution of a discretized problem by linear programming techniques was suggested by Barrodale and Young [14] in 1966 and also by Esch and Eastman in an 1967 technical report (see their 1969 paper [62]). These methods do not make explicit use of characterization results, in contrast to

the (explicit) Remez exchange method of Schumaker presented again in technical reports about the same time (see his 1969 paper [181]). The latter method also solved the discretized problem, but had no convergence results.

Of course any methods for best Chebyshev approximation by linear functions may be used, but a special iterative algorithm for computing best Chebyshev approximations from spline spaces was given by Nürnberger and Sommer [144] in 1983. As in the classical Remes method, a substep at each iteration is the computation of a spline $s \in S_m$ such that

$$(-1)^i(f(\xi_i) - s(\xi_i)) = h, \quad i = 1, \ldots, m+k+2,$$

for some real number h, and given points $\xi_1, \ldots, \xi_{m+k+2}$ in $[a,b]$. The number of equations reflects the fact that S_m has dimension $m+k+1$. Then one of the points ξ_i is replaced by a point where $||f-s||$ is attained in $[a,b]$ to get a new set of points $\{\xi_i\}$. The usual Remes exchange rule can result in a singular system of equations, so a modified exchange rule is needed. Such a rule was given by Nürnberger and Sommer [144], which ensures that the new system has a unique solution. Because of possible nonuniqueness of best approximations, the proof of convergence is fairly complicated. However, convergence can be established.

A multiple exchange procedure can also be implemented, and quadratic convergence is possible. The above results can be extended to more general spline spaces, where the polynomials are replaced by linear combinations of functions forming Chebyshev sets: this was considered by Nürnberger et al. [141] in 1985.

To permit the full power of splines, one should allow the knots to vary, rather than be fixed in advance. The corresponding approximation problem is then a difficult nonlinear problem and we say more about this in Section 3.7.

2.7. Linear L_1 approximation in $C[a,b]$

Given the same setting as at the start of Section 2.5, we consider here the problem

$$\text{find } \boldsymbol{a} \in \mathbb{R}^n \text{ to minimize } \int_a^b \left| f(x) - \sum_{i=1}^n a_i \phi_i(x) \right| \mathrm{d}x. \tag{13}$$

This problem was apparently first considered by Chebyshev in 1889.

Characterization results go back to James [93] in 1947. A convenient form is the analogue of that available in the discrete case: a vector $\boldsymbol{a}$ solves the L_1 problem if and only if there exists a function v with $||v||_\infty \leqslant 1$ such that

$$\int_a^b v(x)\phi_j(x)\,\mathrm{d}x = 0, \quad j = 1, \ldots, n,$$

$$v(x) = \operatorname{sign} r(x), \quad r(x) \neq 0.$$

If the set $\{\phi_1(x), \ldots, \phi_n(x)\}$ forms a Chebyshev set in $[a,b]$, then Jackson [92] in 1921 showed that the solution is unique. For polynomial approximation, perhaps the first "algorithm" was given by Hoel [89] in 1935, who showed that the polynomials of best L_p approximation converge to the best L_1 approximation as $p \to 1$. This is the analogue of the Polya algorithm for Chebyshev approximation. A more general convergence result, and a characterization of the limiting element, was given by Landers and Rogge [109] in 1981.

The L_1 problem is greatly simplified if it can be assumed that the zeros of $f(x) - \sum_{i=1}^{n} a_i \phi_i(x)$ form a set of measure zero in the interval $[a,b]$ (for example the zeros just consist of a finite set of points). Then the function to be minimized in (13) is differentiable, and necessary and sufficient conditions for a solution are that

$$\int_a^b g(x,\boldsymbol{a})\phi_j(x) = 0, \quad j = 1,\ldots,n,$$

where $g(x,\boldsymbol{a})$ denotes the sign of $f(x) - \sum_{i=1}^{n} a_i \phi_i(x)$. This was known to Laasonen [107] in 1949. This means that great store is placed on the points where there are sign changes, or equivalently where the approximation interpolates f. If these points were known, and were exactly n in number, then we could compute the best approximation by interpolation, *provided that there were no other changes of sign in the error of the resulting approximation.* The points $x_1 < \cdots < x_t \in (a,b) = (x_0, x_{t+1})$, where $1 \leqslant t \leqslant n$, are called *canonical points* if

$$\sum_{i=0}^{t} (-1)^i \int_{x_i}^{x_{i+1}} \phi_j(x)\,\mathrm{d}x = 0, \quad j = 1,\ldots,n. \tag{14}$$

For the Chebyshev set case, Laasonen [107] in 1949 showed that there is a unique sign function and further $t = n$. This was extended to weak Chebyshev sets by Micchelli [129] in 1977. Existence of a set of $t \leqslant n$ canonical points for the general problem was shown by Hobby and Rice [87] in 1965.

For the special case when $\phi_i(x) = x^{i-1}$, $i = 1,\ldots,n$, then the location of the n canonical points is known – they lie at the zeroes of the Chebyshev polynomial of the second kind of degree n (shifted if necessary). This result is due to Bernstein [23] in 1926. Thus their location is independent of f. Interpolation at these points can quite frequently result in the best polynomial approximation, for example, if the set

$$\{f(x), \phi_1(x), \ldots, \phi_n(x)\}$$

forms a Chebyshev set in $[a,b]$. However, this is not usually the case, and so this is not a reliable method in general.

An algorithm of descent type seems first to have been given by Usow [188] in 1967, who gave an analysis applicable to problems with Chebyshev sets, and some numerical results for polynomial approximation. However, Marti [120] in 1975 gave an example where the method converges to a point which is not a solution. He gave an alternative descent method, valid when the functions $\{\phi_i\}$ form a Markov set (any rearrangement is a Chebyshev set).

The first general method seems due to Glashoff and Schultz [72] in 1979, based on using Newton's method to solve the characterization conditions (14) together with the corresponding interpolation conditions. A variant of this, which is globally convergent, was given by Watson [200] in 1981. It is essentially of exchange type, based on the calculation of the zeroes of the error at each iteration and the construction of descent directions. It is also of Newton type, since it constructs the Hessian matrix of the error when it exists, and therefore can have a second-order convergence rate. In a sense, it can be thought of as analogous to the second algorithm of Remes for Chebyshev problems, where here a sequence of sets of zeroes plays the role of a sequence of sets of extreme points in that problem; the connection with Newton's method under appropriate circumstances is also something the methods have in common. A method for L_1 problems based on Newton's method was also given by Blatt [25] in 1984.

3. Nonlinear approximation

There are two major differences which arise in moving from linear to nonlinear best approximation problems. Firstly, existence of solutions cannot generally be guaranteed. Secondly, there is normally a gap between conditions which are necessary and conditions which are sufficient for a best approximation. This reflects the loss of convexity. From an algorithmic pont of view, it is usual to seek to satisfy first-order conditions which are necessary for a solution to the best approximation problem, and such a point is conventionally referred to as a *stationary point*. At best this can be expected to be a local minimum of the norm. Assuming that the members of the approximating family are differentiable with respect to the free parameters at least in the region of interest, then a characterization of stationary points is straightforward: it is appropriate simply to replace in the linear case the basis elements (either vectors making up the columns of A or functions $\phi_i, i=1,\ldots,n$) by the partial derivatives of the approximating function with respect to the free parameters at the relevant points.

3.1. Nonlinear approximation in $\mathbb{R}^m$

Consider now the discrete problem

$$\text{find } \boldsymbol{x} \in \mathbb{R}^n \text{ to minimize } ||\boldsymbol{f}(\boldsymbol{x})||,$$

where $\boldsymbol{f} \in \mathbb{R}^m$ depends nonlinearly on the components of $\boldsymbol{x}$, and where the norm is any norm on $\mathbb{R}^m$.

A general approach to this problem is through a sequence of linear subproblems. Assume that $\boldsymbol{f}$ is continuously differentiable in the region of interest, and at a given point $\boldsymbol{x}$, let A denote the $m \times n$ matrix of partial derivatives of the components of $\boldsymbol{f}$ with respect to the components of $\boldsymbol{x}$. Then consider the iterative method based on computing an updated $\boldsymbol{x}$ as follows:

(i) find $\boldsymbol{d} \in \mathbb{R}^n$ to minimize $||\boldsymbol{f} + A\boldsymbol{d}||$,

(ii) replace $\boldsymbol{x}$ by $\boldsymbol{x} + \gamma\boldsymbol{d}$, where $\gamma > 0$ is suitably chosen.

The problem in (i) is just a linear approximation problem in the given norm (a linear subproblem), and (ii) involves choosing γ so that

$$||\boldsymbol{f}(\boldsymbol{x} + \gamma\boldsymbol{d})|| < ||\boldsymbol{f}(\boldsymbol{x})||, \tag{15}$$

if this is possible: for example we may try to minimize the expression on the left-hand side with respect to γ.

When the norm is the least-squares norm, this kind of method (with $\gamma = 1$) most probably dates back to Gauss and is now known as the Gauss–Newton method. For the Chebyshev problem, this kind of approach was suggested by Zuhovickii et al. [212] in 1963, by Ishizaki and Watanabe [91] in 1968, and by Osborne and Watson [150] in 1969. Unless $\boldsymbol{x}$ is a stationary point, improvement can always be obtained via step (ii) since (15) holds for $\gamma > 0$ small enough. The theory given in the Osborne and Watson paper required that successive matrices A satisfied the Haar condition, and in that case convergence to a stationary point was established. The method was extended to the l_1 norm by Osborne and Watson [152] in 1971. Also in 1971, Osborne [146] was able to relax the Haar condition assumption for the l_∞ algorithm, and showed that the method was quadratically convergent if the maximum error at the limit point of the iteration was attained at $n+1$ points. In that case,

unit length steps were ultimately possible, and a ready connection could be drawn with Newton's method applied to the nonlinear equations satisfied at the stationary point. Osborne contrasted this with the behaviour of the method in the l_2 case, when good performance was dependent on the goodness of fit of the model, rather than on properties of the data.

The behaviour of the algorithm in a completely general setting was considered in 1978 by Osborne and Watson [153]; in particular, (15) was always shown to hold for $\gamma > 0$ small enough away from a stationary point. It was also pointed out that the above behaviour typified the situation for polyhedral norms on the one hand, and smooth strictly convex monotonic norms on the other.

A common basis for a convergence analysis which included this kind of algorithm was given by Cromme [51] in 1978: he showed that for second-order convergence, it was sufficient for the best approximation to be strongly unique. This criterion was also studied for the above algorithms in 1980 by Jittorntum and Osborne [95], who showed that strong uniqueness was not always necessary.

Meantime, (at least) two developments were taking place. The fact that the solution of the linear subproblem could be such that very small step lengths were sometimes required led to the idea of explicitly incorporating bounds. This Levenberg–Marquardt or trust region idea was finding favour in descent methods for more general optimization calculations. Another development was to do with the line search. Trying to find the value of γ to minimise $||\boldsymbol{f}||$ is clearly impractical, and the idea of inexact, but sufficiently good, line searches was again imported from contemporary optimization algorithms. These modifications were used by Madsen [115] in an algorithm for the Chebyshev problem, and by Anderson and Osborne [4] in 1977 for polyhedral norm problems (which include l_1 and l_∞). While this could improve things in certain cases, slow convergence could, however, still occur for many problems.

For fast local convergence in general, it was recognized that second derivative information had to be incorporated. Two stage methods for Chebyshev problems were given independently in 1979 by Watson [198] and by Hald and Madsen [80]. These methods solved a sequence of linear subproblems to build up information about the limit point (in particular, the number of points where the norm was attained, with signs). This information could then (if necessary) be used as input to a second (locally convergent) phase such as Newton's method applied to the nonlinear system of equations characterizing a stationary point. Thus they extended fast local convergence to a much wider range of problems.

It had long been recognized that the Chebyshev approximation problem could be posed as a nonlinearly constrained optimization problem, analogous to the way in which the linear problem could, although it seemed at one time that treating the problem in this way was likely to be less efficient than using linear subproblems. However, following advances in techniques for constrained optimization problems, and a recognition that there was much structure in the Chebyshev problem which could be exploited, Conn [48] in 1979, Murray and Overton [135] in 1980, Han [82] in 1981, and Womersley and Fletcher [207] in 1986 all proposed methods. These are all variants of a technique based on the solution of a sequence of quadratic programming problems, involving a Lagrangian function and linearizations of $r_i = h$, for i in a set which estimates the set of indices where the extrema are attained at a stationary point. They all incorporate second derivative information, and involve exploiting the structure and giving descent with respect to the norm. A line search descent method due to Conn and Li [49] in 1989 is claimed to make more explicit use of the structure: in addition to giving descent, it attempts to force satisfaction of the stationary point characterization conditions at the same time.

This general approach now seems the most effective for small problems with dense matrices A. However, for large problems with sparse structure in A, solving linear rather than quadratic programming problems is preferable, as the structure may be exploited. Therefore, for such problems, there has been some recent re-interest in methods of trust region type which use sequential linear programming. Some work of Jonasson and Madsen [97,98] from the mid-1990s is of relevance here.

As in the linear case, large problems may arise as discretizations of continuous problems; therefore we will return to this in Section 3.4.

There were analogous developments for the solution of the nonlinear l_1 problem. The first attempt to incorporate second derivative information into general classes of problems was probably by McLean and Watson [122] in 1980. This method was of two-phase type which used the solution of a sequence of bounded variable linear subproblems to provide information about Z at the desired stationary point, and then used Newton's method to get an accurate point. The exact Jacobian matrix was used for the Newton step. A similar method by Hald and Madsen [81] in 1985 used quasi-Newton approximations, and allowed several switches between phases. Meantime, (single phase) methods based on solving a sequence of quadratic programming problems were being developed, analogous to those mentioned before for Chebyshev approximation problems. In the main, these constructed the quadratic programming problems by defining a Lagrangian function, and by involving linear approximations to $r_i = 0$ for $i \in Z^k$, where Z^k was an estimate at iteration k to Z at the solution. Methods of this type which used line searches were proposed by Murray and Overton [136] in 1981 and Bartels and Conn [15] in 1982, and trust region methods were given by Fletcher [67,68] in 1981 and 1985.

Perhaps because there is no simple connection analogous to that between continuous and discrete Chebyshev approximation problems, the nonlinear l_1 problem has attracted much less recent interest.

3.2. Rational Chebyshev approximation in $\mathbb{R}^t$

Approximation by rational functions goes back to Chebyshev in 1859. The basic (discrete) problem is as follows. Let x_i, $i = 1, \ldots, t$ be in $[a, b]$. Then a best approximation is sought from the set

$$\mathbb{R}^D_{nm} = \left\{ P(x)/Q(x)\colon P(x) = \sum_{j=0}^{n} a_j p_j(x), \quad Q(x) = \sum_{j=0}^{m} b_j q_j(x),\ Q(x_i) > 0,\ i = 1, \ldots, t \right\},$$

to the set of values $f_1, \ldots, f_t$, in the sense that

$$\max_{1 \leqslant i \leqslant t} |\mathbb{R}(x_i) - f_i|$$

is minimized over all $\mathbb{R} \in \mathbb{R}^D_{nm}$. For this problem, existence of best approximations is not guaranteed, even in the case of quotients of polynomials, and characterization and uniqueness results are not available, although of course necessary conditions for a solution may be obtained. In fact necessary conditions based on alternations may be derived analogous to the characterization conditions which are available in the case of approximation to a continuous function on an interval: see Section 3.6. Because of this it is possible to implement an algorithm equivalent to the second algorithm of Remes, although for the discrete problem there are better approaches which do not explicitly use alternations.

The quest for algorithms for rational Chebyshev approximation appears to go back at least as far as Wenzl [201] in 1954. In the late 1950s Loeb considered some approaches which formed the basis for what was perhaps the first really effective algorithm for the discrete problem, the differential correction algorithm, given by Cheney and Loeb [38] in 1961. At that time, the convergence properties were uncertain, and a modified version was subsequently considered by the same authors in 1962 [39], and also by Cheney and Southard [42] in 1963, which was shown to have sure convergence properties, and drew attention away from the original method. However, in 1972 Barrodale et al. [9] studied both approaches, and showed that the method in its original form had not only guaranteed convergence from any starting approximation in $\mathbb{R}^D_{nm}$, but usually had a second-order convergence rate. Further, their comparisons of the methods showed that the performance of the original method was better. Some further analysis was given by Cheney and Powell [41] in 1987.

The differential correction algorithm is an iterative method where successive approximations from $\mathbb{R}^D_{nm}$ are computed by solving a linear programming subproblem, where one variable is minimized subject to $2t$ linear constraints involving also variables representing the coefficients of the new approximation, and bound constraints on the coefficients of the denominator. Each step of the method may be interpreted as working with an approximation to the original problem which is correct up to first order, and this "Newton method" connection gives a partial explanation of the second-order convergence rate. In fact, from the point of view of implementation, it is more efficient to solve the dual of the original linear programming subproblem.

A potentially unsatisfactory feature of approximation from $\mathbb{R}^D_{nm}$ is that the denominator, although positive, can become arbitrarily close to zero at certain points. It is not sufficient simply to impose a lower bound on Q, because of the possibility of multiplying both numerator and denominator by an arbitrary constant. A modification of the differential correction algorithm which applies to problems with a lower bound on the denominator and upper bounds on the absolute values of the coefficients b_j was given by Kaufmann and Taylor [104] in 1981.[5] It is more natural, however, to impose upper and lower bounds on the denominators themselves ("constrained denominators"). A modified differential correction algorithm for this problem has been given by Gugat [75] in 1996. This involves constraints of the form

$$\mu(x_i) \leqslant Q(x_i) \leqslant \nu(x_i), \quad i = 1, \ldots, t, \tag{16}$$

where μ and ν are continuous functions, which replace the constraints on $Q(x_i)$ in the definition of $\mathbb{R}^D_{nm}$.

The linear programming subproblem corresponding to (16) above differs in that the additional conditions are imposed on the denominators. However, Gugat's method differs also in that there is greater flexibility in choice of initial values, and this turns out to be important. The original algorithm starts with an approximation $\mathbb{R}_1$ in $\mathbb{R}^D_{nm}$ and a value Δ_1 which is the maximum modulus error of this approximation on the discrete set. The method of Gugat starts with $\mathbb{R}_1$ as usual, but with an arbitrary number Δ_1 that is allowed to be smaller than the current maximum error. This flexibility turns out to be an important advantage: for example numerical results show that the choice $\Delta_1 = 0$ is a good one. It is shown by Gugat that convergence results for the original version carry over.

[5] This is an example of a constrained problem, which arises in a natural way from the rational approximation problem: it is not our intention to consider constrained problems per se.

It has been pointed out that the quadratic convergence of the differential correction algorithm is a consequence of a connection which it may be shown to have with Newton's method. Methods which set out deliberately to use variants of Newton's method are given by Hettich and Zenke [86] in 1990 and Gugat [76] in 1996. However, in contrast to the methods based on the differential correction algorithm, these do not generate a monotonically decreasing sequence of maximum modulus errors on successive approximations.

3.3. Nonlinear approximation in $C[a,b]$

Consider now the problem

$$\text{find } \boldsymbol{a} \in \mathbb{R}^n \text{ to minimize } ||f(.,\boldsymbol{a})||, \tag{17}$$

where the norm is a given norm on $C[a,b]$ and where $\boldsymbol{a}$ occurs nonlinearly in f. It was shown by Watson [197] in 1978 that, provided that f was differentiable in the region of interest, methods of Gauss–Newton type (the continuous analogues of the methods introduced in Section 3.1) can be applied to this class of problems. However, while this may be of some theoretical interest, it does not lead to practical algorithms. Indeed, such problems cannot really be considered in any generality, and we will in fact restrict attention to the Chebyshev norm, and some important special cases.

3.4. Nonlinear Chebyshev approximation in $C[a,b]$

Here we consider (17) when the norm is the Chebyshev norm

$$||f||_\infty = \max_{x\in[a,b]} |f(x)|.$$

Some general problems of this type were considered by Chebyshev in 1859 [36], with particular reference to rational approximation.

Aside from some special cases (for example see below) it is not possible to say very much about the number of points where the norm is attained at a stationary point. In common with other general nonlinear problems, characterization results are not available, and numerical methods set out to find a stationary point.

The first practical numerical methods seem to have been of two-phase type (see Section 2.5), and these were proposed independently by Hettich [83] and Watson [195] in 1976. The basic idea is similar to that used for linear problems: a first phase is to solve a discretized problem, whose solution identifies the number and associated signs, along with good approximations, of the points where the norm is attained at a stationary point, and a second phase corresponding to the solution of a nonlinear system comprising the equations to be satisfied there. Only the first-phase calculation needs a method which is tailored to whether the problem is linear or not. If a single discretized problem is to be solved, then any of the methods for solving discrete Chebyshev problems can of course be used.

The second phase calculation is a Newton type method, whose steps may be interpreted as quadratic programming problems. The approach can be globalized, thus extending the domain of convergence. This idea was central to the single phase method given by Jonasson and Watson [99] in the mid-1980s, based on the use of a Lagrangian function, and solving a sequence of quadratic programming problems defined on the current set of local maxima of the modulus of the error

function. Descent directions were defined, and both line search and trust region algorithms were developed. Second-order convergence is normal, and there is a nice connection with the second algorithm of Remes; however, the requirement to calculate exact local extrema at each step is a major disadvantage, and there can be sometimes slow progress far from a stationary point. A similar method was given by Jing and Fam [94] in 1987.

The connection between continuous Chebyshev approximation problems and semi-infinite programming problems has already been drawn, and the earlier comments apply to nonlinear problems. It may be that more recent methods being devised for nonlinear semi-infinite programming problems may also improve on these earlier methods for Chebyshev approximation problems. For example, a method by Görner [74] in 1997 consists of the solution of a finite set of discretized problems by sequential quadratic programming methods, following on from similar ideas used by Zhou and Tits [210] in 1996. These methods can lead into a second phase for accurate solution of the continuous problem: a feature of the method of Görner is that the same superlinearly convergent sequential quadratic programming method is used in both phases.

In any event, it would appear that this much at least can be said: a two-phase method with a discretization technique as first phase, and a variant of Newton's method as second phase, seems to be the most reliable and efficient method for solving small to medium size continuous Chebyshev set problems. However, the difficulties referred to near the end of Section 2.5 are still relevant for larger problems.

3.5. Nonlinear Chebyshev approximation in $C[a,b]$ – some special cases

In order to close the gap between conditions which are necessary and conditions which are sufficient, it is necessary to restrict the class of approximating functions, and the point at which this process converges may conveniently be described in terms of alternation conditions, analogous to those which apply in the (linear) Chebyshev set case. This clearly has implications for numerical methods, and so it is appropriate to look briefly at some of this theory. In the linear case, the Chebyshev set condition simultaneously implies the existence of an interpolation function with a certain (fixed) number of zeros. In nonlinear cases, these become two conditions which have to be assumed separately: the interpolation property is a *local* one (which depends on the approximation), but in addition we require a global property on the zeros.

The concept of unisolvency was introduced in 1949 by Motzkin [130]. Let $\phi(x,\boldsymbol{a}):\mathbb{R}^n \to C[a,b]$. Then given any $\boldsymbol{d} \in \mathbb{R}^n$ and n distinct points x_i, $i=1,\ldots,n$ in $[a,b]$, this family is *unisolvent* if there exists a unique vector $\boldsymbol{a} \in \mathbb{R}^n$ such that

$$\phi(x_i,\boldsymbol{a}) = d_i, \quad i=1,\ldots,n.$$

This particular generalization of the Chebyshev set property in the linear case leads to the existence of a unique best approximation ϕ which is characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \geqslant n+1,$$

as shown by Tornheim [187] in 1950. Unfortunately this is an extremely restrictive property, possessed by a small number of approximating functions, and Rice in his Ph.D. thesis in 1959, and in papers published in the next few years, suggested a more general property of *varisolvency*, which

(provided the error is not constant) leads to the best approximation ϕ being characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \geqslant m(\phi)+1,$$

where $m(\phi)$ is the degree of local solvency [168,170]. Rice also showed that there is at most one best approximation. If ϕ is formed from a linear combination of n functions forming a Chebyshev set in $[a,b]$, then this is in fact a varisolvent family of constant degree n.

A related theory for nonlinear Chebyshev approximation on an interval was established by Meinardus and Schwedt [126] in 1964, valid for approximating functions differentiable with respect to their parameters. It essentially replaces the local condition required in varisolvency by a local Chebyshev set condition on the tangent space. An alternation characterization condition was established, along with an uniqueness result. Braess [31] in 1974 demonstrated the precise relationship between these various results.

Attempting to define a general class of nonlinear approximating functions which would be varisolvent, and so satisfy this kind of characterization result, Hobby and Rice [88] in 1967 defined *γ-polynomials*,

$$\phi(x,\boldsymbol{a}) = \sum_{i=1}^{n} a_i \gamma(a_{i+n}x),$$

where γ is a continuous function of its parameters. This class is of interest because it includes some important special cases, for example exponentials and spline functions. Subject to an additional assumption (Descartes' rule of signs), Hobby and Rice [88] established that the theory of varisolvent families applied. This condition is satisfied if the set

$$\{\gamma(t_1,x),\ldots,\gamma(t_n,x)\}$$

forms a Chebyshev set in $[a,b]$ for distinct t_i's. A best approximation ϕ is then characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \geqslant n+l(\phi)+1$$

where $l(\phi)$ is the length of the γ-polynomial ϕ, defined by the restriction that ϕ cannot be expressed by a sum of fewer terms. The closure of the set of γ-polynomials is in fact required for existence of best approximations, but then the alternating characterization is lost.

An important special case is given by taking

$$\gamma(t,x) = \mathrm{e}^{tx},$$

when we have approximation by sums of exponentials. This was studied first by Rice [169] in 1960 ($n=1$), and in 1962 (general n) [171]. Because the set $\{\mathrm{e}^{t_1x},\ldots,\mathrm{e}^{t_nx}\}$ forms a Chebyshev set in $[a,b]$ for distinct t_i's, then a Descartes' rule of signs holds (this result seems to go back to Laguerre [108] in 1898), and it follows that the approximating family is varisolvent. This was shown by Rice [171] in 1962, who also showed that a best approximation ϕ is characterized by

$$\mathscr{A}(f-\phi)_{[a,b]} \geqslant n+k(\phi)+1,$$

where the gradient vector of ϕ with respect to a_i, $i=1,\ldots,2n$ has $n+k(\phi)$ nonzero components. There is at most one best approximation. Existence of best approximations from the closure of the set of exponential functions was proved by Rice [171] in 1962 and Werner [204,205] in 1969.

As Bellman [21] wrote in 1970, "exponential approximation is a notoriously delicate enterprise", mainly because widely varying parameter values can give nearly optimal results. Therefore, the calculation of best Chebyshev approximations (or indeed any approximations) by sums of exponentials can be difficult. If an assumption is made about the number of alternations (that $k(\phi) = n$), then a method of Remes type can be applied with a nonlinear system of equations to be solved for the new coefficients at each iteration. This is considered by Dunham [57] in 1979, and in subsequent work with Zhu: it was necessary to have very good starting approximations.

The fact that n of the parameters occur linearly means that if the parameters $a_{n+1}, \ldots, a_{2n}$ (the frequencies) are fixed, then the remaining parameters can be obtained by applying a linear solution method; this gives a problem which is essentially in the frequencies alone, and which could be tackled by iteration on the frequencies to obtain optimal values. Local descent methods were suggested by Braess [29] and Werner [205] in the late 1960s, and related methods were implemented in the 1970s by others such as Cromme, Kammler, Robitzsch and Schaback [50,101,177]. A method due to Dunham [59] in 1988 worked well with one frequency, but had difficulties with two or more. Nearly equal frequencies, or coalescing frequencies, are generally a problem.

One feature is that good initial approximations are necessary: in particular it is important to estimate the positions of the frequencies, before applying an optimal method, and this has led to interest in "suboptimal approximations". Prony's method of "approximate interpolation" may be applied, although the method is not generally stable. An alternative is Bellman's 1970 [21] method of differential approximation. These methods were considered in detail by Robitzsch and Schaback [177] in 1978 and by Schaback [179] in 1979. Any suboptimal method may be considered as a first phase method which can lead into a second phase based on Newton's method to satisfy the nonlinear system characterizing the solution.

But it would seem that in practice additional constaints are both natural physically, and necessary mathematically and computationally – for example, to bound frequencies, or to prevent frequencies from crossing each other. The computational approach then depends on precisely what is being assumed, and we will not pursue this further.

3.6. Rational Chebyshev approximation in $C[a,b]$

The continuous analogue of the class of problems considered in Section 3.2 is based on the approximating set $\mathbb{R}_{nm}$ defined by

$$\mathbb{R}_{nm} = \left\{ P(x)/Q(x) : P(x) = \sum_{j=0}^{n} a_j p_j(x), \right.$$
$$\left. Q(x) = \sum_{j=0}^{m} b_j q_j(x),\ Q(x) > 0 \text{ on } [a,b] \right\},$$

where the $p_j(x)$ and $q_j(x)$ are given sets of functions. Then given $f(x) \in C[a,b]$, we require to determine $\mathbb{R} \in \mathbb{R}_{nm}$ to minimize $||f - \mathbb{R}||$, where the norm is the Chebyshev norm on $[a,b]$. For the special case when $P(x)$ and $Q(x)$ are polynomials of degree n and m, respectively, existence of a best approximation is guaranteed, as shown by Walsh [193] in 1931. Achieser in 1947 (see his 1956 book [3]) showed that the best approximation is unique (up to a normalization), and earlier,

in 1930 (again see his 1956 book [3]), he showed that a best approximation $\mathbb{R} = P/Q \in \mathbb{R}_{nm}$ is characterized by

$$\mathscr{A}(f - \mathbb{R})_{[a,b]} \geqslant n + m + 2 - d(\mathbb{R}),$$

where $d(\mathbb{R})$ is the *defect* of the approximation: the defect is just the minimum difference between the *actual* degree of $P(x)$ and $Q(x)$ and n and m respectively. If $d(\mathbb{R}) > 0$, the best approximation is said to be degenerate. These results also follow from the fact that the approximating family is varisolvent. (The necessary conditions referred to in Section 3.2 correspond to this alternation result defined *on the points of the set* $x_1 < \cdots < x_t$ introduced there.)

For more general quotients (of linear combinations of functions), existence is no longer guaranteed, although characterization results are available (not necessarily of alternation type), and uniqueness results may be extended. The main contributions here are from Cheney and Loeb [39,40] in the mid-1960s.

For rational approximation by quotients of polynomials on an interval, the analogue of the Remes exchange method may be applied, using sets of $m+n+2$ points. It, therefore, requires nondegeneracy of the best approximation, and can converge at a second-order rate if started from close enough to the solution: the analysis is primarily due to Werner in a series of papers in the early 1960s [202,203]. The system of linear equations which needs to be solved in the linear problem is replaced by a nonlinear system in the rational problem, equivalent to an eigenvalue problem. Werner [203] in 1963 showed that the eigenvalues are always real, and there is at most one pole free solution, that is a rational approximation with $Q(x) > 0$ on $[a, b]$. Maehly in 1963 [118] gave an example which showed that in fact no pole free solution need exist; even if it is does exist, it need not be associated with the smallest eigenvalue. Despite these potential problems, the second algorithm of Remes has been successfully used for rational approximation. Fraser and Hart [70] in 1962, Werner [202] in 1962 and Stoer [186] in 1964 gave methods based on solving the system of nonlinear equations directly. In 1966, Curtis and Osborne [52] gave an algorithm which used the eigenvalue connection explicitly, solving the eigenvalue problem by inverse iteration with zero as an initial estimate for the eigenvalue; they also established quadratic convergence. Breuer [32] in 1987 suggested a different direct approach to this subproblem which used continued fraction interpolation, and which it was claimed can lead to a considerable increase in efficiency, and also accuracy and robustness.

Variants of the second algorithm of Remes apply to rational Chebyshev approximation problems which incorporate a generalized weight function. Important work involving rational approximation on an interval to provide optimal starting values for computing $\sqrt{x}$ by the Newton Raphson method was done, for example, by Moursand [133] in the late 1960s.

The algorithms fail if the solution is degenerate, and indeed for problems which are nearly degenerate, extremely good starting approximations are required. Ralston [160,161] in a series of papers in the late 1960s and early 1970s considered degeneracy in detail. The computation of nearly degenerate approximations should if possible be avoided, as equally good results can be obtained through the use of smaller m and n.

It is possible to make the second algorithm of Remes more robust, by combining its merits with the differential correction algorithm. In particular the discrete subproblems can be solved by that method, and if no pole-free solution is obtained, additional points can be included. If sufficiently many points are taken in $[a, b]$, and always assuming that the continuous problem is not degenerate, then a pole-free solution can be obtained so that the algorithm can be continued. Methods based

on this idea were given by Belogus and Liron [22] and also Kaufman et al. [103] both in 1978. Numerical evidence is that such an approach can be successful for problems which give difficulties with the traditional Remes method.

The differential correction algorithm may be applied to problems defined on an interval, although the subproblems are no longer finite. Dua and Loeb [55] in 1973 established a second order convergence rate if the best approximation is normal. The potentially unsatisfactory feature referred to in Section 3.2 where the deniminator, although positive, can become arbitrarily close to zero, also applies to $\mathbb{R}_{nm}$. The algorithm of Gugat referred to there also may be applied to intervals, although the numerical performance is unclear.

3.7. Chebyshev approximation by spline functions with free knots

To permit the full power of splines, one should allow the knots to vary, rather than be fixed in advance. The corresponding approximation problem is then a difficult nonlinear problem. This problem can be considered in terms of γ polynomials. However, the structure of the problem, and the way in which degeneracies can be introduced makes an attempt to make a straightforward application unhelpful.

To guarantee existence of best approximations, multiple knots have to be allowed. There may be local solutions; a characterization of best approximations is not known. For the case of k free knots, necessary and (different) sufficient conditions of the alternation kind given above may be proved. Let q' denote the sum of the knot multiplicities at the points $x_{p+1},\ldots,x_{p+q-1}$. Then it is *necessary* for $s \in S_m$ to be a best Chebyshev approximation with k free knots to f in $[a,b]$ that there exists an interval $[x_p,x_{p+q}] \subset [a,b]$ with $q \geqslant 1$ such that

$$\mathscr{A}(f-s)_{[x_p,x_{p+q}]} \geqslant m+q+q'+1,$$

as shown by Nürnberger et al. [142] in 1989; it is *sufficient* for $s \in S_m$ to be a best Chebyshev approximation with k free knots to f in $[a,b]$ that there exists an interval $[x_p,x_{p+q}] \subset [a,b]$ with $q \geqslant 1$ such that

$$\mathscr{A}(f-s)_{[x_p,x_{p+q}]} \geqslant m+k+q'+2,$$

as shown by Braess [30] in 1971. The necessary condition was strengthened to a possibly longer alternant by Mulansky [134] in 1992. Although a characterization of best spline approximations with free knots is not known, a characterization of strongly unique best spline approximations with free simple knots is available: what is required is that *all* knot intervals contain sufficiently many alternating extrema. The relevant work here is by Nürnberger [137,138] in 1987 and 1994.

Since approximation by splines with free knots is a nonlinear Chebyshev approximation problem, of course general methods can be used. However, the way in which the knots enter as free parameters makes this a particularly awkward problem and makes it important that the special structure be exploited.

For a discretization of the problem, a descent method based on Newton's method was given by Esch and Eastman [62] in 1969. Most algorithmic work has been concerned with uniform approximation from a space of piecewise polynomials where the continuity conditions at the knots are relaxed. A standard algorithmic approach is based on so-called segment approximation, originating from work of Lawson in 1964 [112], and methods were proposed by Pavlidis and Maika [155] in

1974, and McLaughlin and Zacharski [121] in 1984. Because pieces were fitted separately, continuity could be lost between segments. A recent method of this type is due to Nürnberger et al. [145] in 1986 (see also [124]). The algorithm converges through sequences of knot sets from an arbitrary set of knots. For each set of k knots, best Chebyshev degree m polynomial approximations to f are obtained on each subinterval using the classical Remes algorithm. The knots are then adjusted by a "levelling" process, so that the maximum errors of the polynomial best approximations are equalized. The result of this is a piecewise polynomial which is usually discontinuous. However, the procedure is augmented by the application of the method for fitting splines with fixed knots to the optimal knot positions obtained from the first part. The outcome of this is a differentiable spline approximation, which numerical results show to be a good one. Note that at present there is no algorithm for computing (global) best Chebyshev spline approximations with free knots. At best a local approximation can be expected, so producing a "good" spline approximation may be the most sensible strategy.

Generalizations to multivariate splines have mainly been concerned with interpolation problems. But consider bivariate splines on $[a_1, b_1] \times [a_2, b_2]$. This region can be divided into rectangles by knot lines $x = x_i,\ y = y_i,\ i = 1, \ldots, s$, and a tensor product spline space can be defined. As in the univariate problem, partitions can be defined and improved systematically in such a way that best Chebyshev approximations are obtained in the limit. Some recent work on this problem is given by Meinardus et al. [125] in 1996, and by Nürnberger [140] in 1997. However, there are many unsolved problems, as pointed out by Nürnberger [139] in 1996.

Acknowledgements

Charles Dunham attributes to P. Whippey, on the effort of writing a history: "It's a complex problem: the costs are real, the benefits imaginary". In any event, I am grateful to the many people who took the trouble to answer my questions about aspects of this work, or provided me with references. I am especially grateful to Mike Osborne, because he read through a draft and provided me with many helpful comments and suggestions. Of course, the responsibility for what is set down here is entirely mine.

Mike Osborne in fact has to shoulder quite a lot of responsibility for the existence of this paper, because he introduced me to the subject of approximation in 1964, he supervised my postgraduate study, and generally he has remained since that time a strong and guiding influence on my career. I dedicate this work to him, on the occasion of his 65th birthday, with my gratitude and my affection.

References

[1] N.N. Abdelmalik, Linear L_1 approximation for a discrete point set and L_1 solutions of overdetermined linear equations, J. Assoc. Comput. Mach. 18 (1971) 41–47.

[2] N.N. Abdelmalik, An efficient method for the discrete linear L_1 approximation problem, Math. Comp. 29 (1975) 844–850.

[3] N.I. Achieser, Theory of Approximation, Ungar, New York, 1956.

[4] D.H. Anderson, M.R. Osborne, Discrete, nonlinear approximation problems in polyhedral norms, Numer. Math. 28 (1977) 143–156.

[5] D.O. Andreassen, G.A. Watson, Linear Chebyshev approximation without Chebyshev sets, BIT 16 (1976) 349–362.

[6] R.D. Armstrong, J. Godfrey, Two linear programming algorithms for the linear discrete L_1 norm problem, Math. Comp. 33 (1979) 289–300.
[7] R.D. Armstrong, D.S. Kung, A dual method for discrete Chebyshev curve fitting, Math. Programming 19 (1979) 186–199.
[8] I. Barrodale, C. Phillips, Algorithm 495: solution of an overdetermined system of linear equations in the Chebyshev norm, ACM Trans. Math. Software 1 (1975) 264–270.
[9] I. Barrodale, M.J.D. Powell, F.D.K. Roberts, The differential correction algorithm for rational l_∞ approximation, SIAM J. Numer. Anal. 9 (1972) 493–504.
[10] I. Barrodale, F.D.K. Roberts, Applications of mathematical programming to l_p approximation, in: J.B. Rosen, O.L. Mangasarian, K. Ritter (Eds.), Nonlinear Programming, Academic Press, New York, 1970, pp. 447–464.
[11] I. Barrodale, F.D.K. Roberts, An improved algorithm for discrete l_1 linear approximation, SIAM J. Numer. Anal. 10 (1973) 839–848.
[12] I. Barrodale, F.D.K. Roberts, Algorithm 478: solution of an overdetermined system of equations in the L_1 norm, Comm. ACM 17 (1974) 319–320.
[13] I. Barrodale, A. Young, Algorithms for best L_1 and L_∞ linear approximation on a discrete set, Numer. Math. 8 (1966) 295–306.
[14] I. Barrodale, A. Young, A note on numerical procedures for approximation by spline functions, Comput. J. 9 (1966) 318–320.
[15] R. Bartels, A.R. Conn, An approach to nonlinear l_1 data fitting, in: J.P. Hennart (Ed.), Numerical Analysis, Cocoyoc, 1981, Springer, Berlin, 1982, pp. 48–58.
[16] R. Bartels, A.R. Conn, C. Charalambous, On Cline's direct method for solving overdetermined linear systems in the l_∞ sense, SIAM J. Numer. Anal. 15 (1978) 255–270.
[17] R. Bartels, A.R. Conn, Y. Li, Primal methods are better than dual methods for solving overdetermined linear systems in the l_∞ sense, SIAM J. Numer. Anal. 26 (1989) 693–726.
[18] R. Bartels, A.R. Conn, J.W. Sinclair, Minimization techniques for piecewise differentiable functions: the l_1 solution to an overdetermined linear system, SIAM J. Numer. Anal. 15 (1978) 224–241.
[19] R. Bartels, G.H. Golub, Stable numerical methods for obtaining the Chebyshev solution to an overdetermined system of equations, Comm. ACM. 11 (1968) 401–406.
[20] A.E. Beaton, J.W. Tukey, The fitting of power series, meaning polynomials, illustrated on band-spectrographic data, Technometrics 16 (1974) 147–185.
[21] R. Bellmann, Methods of Nonlinear Analysis I, Academic Press, New York, 1970.
[22] D. Belogus, N. Liron, DCR2: an improved algorithm for l_∞ rational approximation, Numer. Math. 31 (1978) 17–29.
[23] S.N. Bernstein, Lecons sur les Propriétés Extrémales et la Meillure Approximation des Fonctions Analytiques d'une Variable Réelle, Gauthier-Villars, Paris, 1926.
[24] L. Bittner, Das Austauschverfahren der linearen Tschebyscheff-Approximation bei nicht erfullter Haarscher Bedingung, Z. Angew. Math. Mech. 41 (1961) 238–256.
[25] H.-P. Blatt, Exchange algorithms, error estimates and strong unicity in convex programming and Chebyshev approximation, in: S.P. Singh, J.W.H. Burry, B. Watson (Eds.), Approximation Theory and Spline Functions, Reidel, Dodrecht, 1984, pp. 1–41.
[26] P. Bloomfield, W.L. Steiger, Least Absolute Deviations, Birkhaüser, Boston, 1983.
[27] P.T. Boggs, A new algorithm for the Chebyshev solution of overdetermined linear systems, Math. Comp. 28 (1974) 203–217.
[28] E. Borel, Lecons sur les Fonctions de Variables Reélles, Gauthier-Villars, Paris, 1905.
[29] D. Braess, Approximation mit Exponentialsummen, Computing 2 (1967) 309–321.
[30] D. Braess, Chebyshev approximation by spline functions with free knots, Numer. Math. 17 (1971) 357–366.
[31] D. Braess, Geometrical characterization for nonlinear uniform approximation, J. Approx. Theory 11 (1974) 260–274.
[32] P.T. Breuer, A new method for real rational uniform approximation, in: J.C. Mason, M.G. Cox (Eds.), Algorithms for Approximation, Clarendon Press, Oxford, 1987, pp. 265–283.
[33] P. Butzer, F. Jongmans, P.L. Chebyshev (1821–1894). A guide to his life and work, J. Approx. Theory 96 (1999) 11–138.

[34] C. Carasso, P.J. Laurent, An algorithm of successive minimization in convex programming, R. A. I. R. O. Numer. Anal. 12 (1978) 377–400.
[35] A. Charnes, W.W. Cooper, R.O. Ferguson, Optimal estimation of executive compensation by linear programming, Management Sci. 1 (1955) 138–151.
[36] P.L. Chebyshev, Sur les questions de minima qui se rattachent à la representation approximative des fonctions, Oeuvres I (1859) 273–378.
[37] E.W. Cheney, A.A. Goldstein, Newton's method for convex programming and Tchebycheff approximation, Numer. Math. 1 (1959) 253–268.
[38] E.W. Cheney, H.L. Loeb, Two new algorithms for rational approximation, Numer. Math. 3 (1961) 72–75.
[39] E.W. Cheney, H.L. Loeb, On rational Chebyshev approximation, Num. Math. 4 (1962) 124–127.
[40] E.W. Cheney, H.L. Loeb, Generalized rational approximation, SIAM J. on Num. Anal. 1 (1964) 11–25.
[41] E.W. Cheney, M.J.D. Powell, The differential correction algorithm for generalized rational functions, Constr. Approx. 3 (1987) 249–256.
[42] E.W. Cheney, T.H. Southard, A survey of methods for rational approximation, with particular reference to a new method based on a formula of Darboux, SIAM Rev. 5 (1963) 219–231.
[43] J.F. Claerbout, F. Muir, Robust modelling with erratic data, Geophysics 38 (1973) 826–844.
[44] D.I. Clark, The mathematical structure of Huber's M-estimator, SIAM J. Sci. Statist. Comput. 6 (1985) 209–219.
[45] A.K. Cline, A descent method for the uniform solution to overdetermined systems of linear equations, SIAM J. Numer. Anal. 13 (1976) 293–309.
[46] T.F. Coleman, Y. Li, A globally and quadratically convergent affine scaling algorithm for l_1 problems, Math. Programming 56 (1992) 189–222.
[47] T.F. Coleman, Y. Li, A globally and quadratically convergent method for linear l_∞ problems, SIAM J. Numer. Anal. 29 (1992) 1166–1186.
[48] A.R. Conn, An efficient second order method to solve the (constrained) minimax problem, Report CORR 79-5, University of Waterloo, 1979.
[49] A.R. Conn, Y. Li, An efficient algorithm for nonlinear minimax problems, Research Report CS-88-41, University of Waterloo, 1989.
[50] L. Cromme, Eine Klasse von Verfahren zur Ermittlung bester nichtlinearen Tschebyscheff-Approximationen, Numer. Math. 25 (1976) 447–459.
[51] L. Cromme, Strong uniqueness: a far reaching criterion for the convergence of iterative processes, Numer. Math. 29 (1978) 179–193.
[52] A.R. Curtis, M.R. Osborne, The construction of minimax rational approximations to functions, Comput. J. 9 (1966) 286–293.
[53] M. Davies, Linear approximation using the criterion of least total deviations, J. Roy. Statist. Soc. Ser. B 29 (1967) 101–109.
[54] J. Descloux, Approximations in L^p and Chebyshev approximations, SIAM J. 11 (1963) 1017–1026.
[55] S.N. Dua, H.L. Loeb, Further remarks on the differential correction algorithm, SIAM J. Numer. Anal. 10 (1973) 123–126.
[56] A.M. Duarte, R.J. Vanderbei, Interior point algorithms for lsad and lmad estimation, Technical Report SOR-94-07, Programs in Operations Research and Statistics, Princeton University, 1994.
[57] C.B. Dunham, Chebyshev approximation by exponential-polynomial sums, J. Comput. Appl. Math. 5 (1979) 53–57.
[58] C.B. Dunham, The weakened first algorithm of Remes, J. Approx. Theory 31 (1981) 97–98.
[59] C.B. Dunham, Approximation with one (or few) parameters nonlinear, J. Comput. Appl. Math. 21 (1988) 115–118.
[60] R. Dutter, Robust regression: different approaches to numerical solutions and algorithms, Research Report No 6, Technische Hochschule, Zurich, 1975.
[61] F.Y. Edgeworth, On a new method of reducing observations relating to several quantities, Philos. Mag. 5 (1888) 185–191.
[62] R.E. Esch, W.L. Eastman, Computational methods for best spline function approximation, J. Approx. Theory 2 (1969) 85–96.
[63] R.W. Farebrother, The historical development of the L_1 and L_∞ estimation procedures 1793–1930, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amsterdam, 1987, pp. 37–63.

[64] R.W. Farebrother, Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900, Springer, Berlin, 1999.
[65] J. Fischer, The convergence of the best discrete linear L_p approximation as $p \to 1$, J. Approx. Theory 39 (1983) 374–385.
[66] W.D. Fisher, A note on curve fitting with minimum deviations by linear programming, J. Amer. Statist. Assoc. 56 (1961) 359–362.
[67] R. Fletcher, Numerical experiments with an exact l_1 penalty function method, in: O.L. Mangasarian, R.R. Meyer, S.M. Robinson (Eds.), Nonlinear Programming 4, Academic Press, New York, 1981, pp. 99–129.
[68] R. Fletcher, An l_1 penalty method for nonlinear constraints, in: P.T. Boggs, R.H. Byrd, R.B. Schnabel (Eds.), Numerical Optimization 1984, SIAM Publications, Philadelphia, 1985, pp. 26–40.
[69] R. Fletcher, J.A. Grant, M.D. Hebden, Minimax approximation as the limit of best L_p approximation, SIAM J. Numer. Anal. 11 (1974) 123–136.
[70] W. Fraser, J.F. Hart, On the computation of rational approximations to continuous functions, Comm. ACM 5 (1962) 401–403.
[71] K. George, M.R. Osborne, The efficient computation of linear rank statistics, J. Comput. Simul. 35 (1990) 227–237.
[72] K. Glashoff, R. Schultz, Uber die genaue Berechnung von besten L^1-Approximierenden, J. Approx. Theory 25 (1979) 280–293.
[73] G.H. Golub, L.B. Smith, Algorithm 414: Chebyshev approximation of continuous functions by a Chebyshev system, Comm. ACM 14 (1971) 737–746.
[74] S. Görner, Ein Hybridverfahren zur Losung nichtlinearer semi-infiniter Optimierungsprobleme, Ph.D. Thesis, Technical University of Berlin, 1997.
[75] M. Gugat, An algorithm for Chebyshev approximation by rationals with constrained denominators, Constr. Approx. 12 (1996) 197–221.
[76] M. Gugat, The Newton differential correction algorithm for uniform rational approximation with constrained denominators, Numer. Algorithms 13 (1996) 107–122.
[77] S.-A. Gustafson, On the computational solution of a class of generalized moment problems, SIAM J. Numer. Anal. 7 (1970) 343–357.
[78] S.-A. Gustafson, K. Kortanek, Numerical treatment of a class of semi-infinite programming problems, Naval. Res. Logist. Quart. 20 (1973) 477–504.
[79] A. Haar, Die Minkowskische Geometrie und die Annaherung an stetige Funktionen, Math. Ann. 78 (1918) 294–311.
[80] J. Hald, K. Madsen, A two-stage algorithm for minimax optimization, in: A. Bensoussan, J. Lions (Eds.), International Symposium on Systems Optimization and Analysis, Springer, Berlin, 1979, pp. 225–239.
[81] J. Hald, K. Madsen, Combined LP and quasi-Newton methods for nonlinear l_1 optimization, SIAM J. Numer. Anal. 22 (1985) 68–80.
[82] S.-P. Han, Variable metric methods for minimizing a class of nondifferentiable functions, Math. Prog. 20 (1981) 1–13.
[83] R. Hettich, A Newton method for nonlinear Chebyshev approximation, in: R. Schaback, K. Scherer (Eds.), Approximation Theory, Springer, Berlin, 1976, pp. 222–236.
[84] R. Hettich, Numerical methods for nonlinear Chebyshev approximation in: G. Meinardus (Ed.), Approximation in Theorie und Praxis, B.I.-Wissenschaftsverlag, Mannheim, 1979, pp. 139–156.
[85] R. Hettich, An implementation of a discretization method for semi-infinite programming, Math. Programming 34 (1986) 354–361.
[86] R. Hettich, P. Zenke, An algorithm for general restricted rational Chebyshev approximation, SIAM J. Numer. Anal. 27 (1990) 1024–1033.
[87] C.R. Hobby, J.R. Rice, A moment problem in L_1-approximation, Proc. Amer. Math. Soc. 65 (1965) 665–670.
[88] C.R. Hobby, J.R. Rice, Approximation from a curve of functions, Arch. Rational Mech. Anal. 27 (1967) 91–106.
[89] P.G. Hoel, Certain problems in the theory of closest approximation, Amer. J. Math. 57 (1935) 891–901.
[90] M.J. Hopper, M.J.D. Powell, A technique that gains speed and accuracy in the minimax solution of overdetermined linear equations, in: J.R. Rice (Ed.), Mathematical Software III, Academic Press, New York, 1977.
[91] Y. Ishizaki, H. Watanabe, An iterative Chebyshev approximation method for network design, IEEE Trans. Circuit Theory 15 (1968) 326–336.
[92] D. Jackson, Note on a class of polynomials of approximation, Trans. Amer. Math. Soc. 13 (1921) 320–326.

[93] R.C. James, Orthogonality and linear functionals in normed linear spaces, Trans. Amer. Math. Soc. 61 (1947) 265–292.
[94] Z. Jing, A.T. Fam, An algorithm for computing continuous Chebyshev approximations, Math. Comp. 48 (1987) 691–710.
[95] K. Jittorntrum, M.R. Osborne, Strong uniqueness and second order convergence in nonlinear discrete approximation, Numer. Math. 34 (1980) 439–455.
[96] R.S. Johnson, On monosplines of least deviation, Trans. Amer. Math. Soc. 96 (1960) 458–477.
[97] K. Jonasson, A projected conjugate gradient method for sparse minimax problems, Numer. Algorithms 5 (1993) 309–323.
[98] K. Jonasson, K. Madsen, Corrected sequential linear programming for sparse minimax optimization, BIT 34 (1994) 372–387.
[99] K. Jonasson, G.A. Watson, A Lagrangian method for multivariate continuous Chebyshev approximation problems, in: W. Schempp, K. Zeller (Eds.), Multivariate Approximation Theory 2, I.S.N.M. 61, Birkhaüser, Basel, 1982, pp. 211–221.
[100] R.C. Jones, L.A. Karlovitz, Equioscillation under nonuniqueness in the approximation of continuous functions, J. Approx. Theory 3 (1970) 138–145.
[101] D.W. Kammler, Chebyshev approximation of completely monotonic functions by sums of exponentials, SIAM J. Numer. Anal. 13 (1976) 761–774.
[102] N. Karmarker, A new polynomial time algorithm for linear programming, Combinatorica 4 (1984) 373–395.
[103] E.H. Kaufmann, D.J. Leeming, G.D. Taylor, A combined Remes-Differential Correction algorithm for rational approximation, Math. Comp. 32 (1978) 233–242.
[104] E.H. Kaufmann, G.D. Taylor, Uniform approximation by rational functions having restricted denominators, J. Approx. Theory 32 (1981) 9–26.
[105] J.E. Kelley Jr., An application of linear programming to curve fitting, SIAM J. 6 (1958) 15–22.
[106] P. Kirchberger, Uber Tchebychefsche Annaherungsmethoden, Math. Ann. 57 (1903) 509–540.
[107] P. Laasonen, Einige Satze uber Tschebyscheffsche Funktionensysteme, Ann. Acad. Sci. Fenn. Ser. AI 52 (1949) 3–24.
[108] E. Laguerre, Ouvres I, Gauthier-Villars, Paris, 1898.
[109] D. Landers, L. Rogge, Natural choice of L_1-Approximants, J. Approx. Theory 33 (1981) 268–280.
[110] P.S. Laplace, Mechanique Celeste, Tome 111, No 39, 1799.
[111] C.L. Lawson, Contributions to the Theory of Linear Least Maximum Approximation, Ph.D. Dissertation, University of California, Los Angeles, 1961.
[112] C.L. Lawson, Characteristic properties of the segmented rational minimax approximation problem, Num. Math. 6 (1964) 293–301.
[113] W. Li, J.J. Swetits, The linear l_1 estimator and Huber M-estimator, SIAM J. on Optim. 8 (1998) 457–475.
[114] Y. Li, A globally convergent method for L_p problems, SIAM J. Optim. 3 (1993) 609–629.
[115] K. Madsen, An algorithm for minimax solution of over-determined systems of non-linear equations, J. Inst. Math. Appl. 16 (1975) 321–328.
[116] K. Madsen, H.B. Neilsen, A finite smoothing algorithm for linear l_1 estimation, SIAM J. Optim. 3 (1993) 223–235.
[117] K. Madsen, H.B. Nielson, M.C. Pinar, New characterizations of l_1 solutions of overdetermined linear systems, Operations Research Letters 16 (1994) 159–166.
[118] H.J. Maehly, Methods of fitting rational approximations, Part II, J. Assoc. Comput. Mach. 10 (1963) 257–266.
[119] J.C. Mairhuber, On Haar's theorem concerning Chebyshev approximation problems having unique solutions, Proc. Amer. Math. Soc. 7 (1956) 609–615.
[120] J.T. Marti, A method for the numerical computation of best L_1 approximations of continuous functions, Proceedings of Oberwolfach Conference, ISNM, Vol. 26, Birkhauser, Basel, 1975.
[121] H.W. McLaughlin, J.J. Zacharski, Segmented approximation, in: E.W. Cheney (Ed.), Approximation Theory III, Academic Press, New York, 1980, pp. 647–654.
[122] R.A. McLean, G.A. Watson, Numerical methods for nonlinear discrete L_1 approximation problems, Proceedings of Oberwolfach Conference on Numerical Methods in Approximation Theory 1979, ISNM, Vol. 52, Birkhauser, Basel, 1980.
[123] S. Mehrotra, On the implementation of a primal-dual interior point method, SIAM J. Optim. 2 (1992) 575–601.

[124] G. Meinardus, G. Nürnberger, M. Sommer, H. Strauss, Algorithms for piecewise polynomials and splines with free knots, Math. Comp. 53 (1989) 235–247.
[125] G. Meinardus, G. Nürnberger, G. Walz, Bivariate segment approximation and splines, Adv. Comput. Math. 6 (1996) 25–45.
[126] G. Meinardus, D. Schwedt, Nicht-lineare Approximationen, Arch. Rational Mech. Anal. 17 (1964) 297–326.
[127] M.S. Meketon, Least absolute value regression, Technical Report, AT&T Bell Laboratories, Murray Hill, New Jersey, 1987.
[128] G. Merle, H. Späth, Computational experience with discrete L_p approximation, Computing 12 (1974) 315–321.
[129] C.A. Micchelli, Best L^1 approximation by weak Chebyshev systems and the uniqueness of interpolating perfect splines, J. Approx. Theory 19 (1977) 1–14.
[130] T.S. Motzkin, Approximation by curves of a unisolvent family, Bull. Amer. Math. Soc. 55 (1949) 789–793.
[131] T.S. Motzkin, J.L. Walsh, Least pth power polynomials on a real finite point set, Trans. Amer. Math. Soc. 78 (1955) 67–81.
[132] T.S. Motzkin, J.L. Walsh, Least pth power polynomials on a finite point set, Trans. Amer. Math. Soc. 83 (1956) 371–396.
[133] D.G. Moursund, Optimal starting values for Newton-Raphson calculation of $\sqrt{x}$, Comm. ACM 10 (1967) 430–432.
[134] B. Mulansky, Chebyshev approximation by spline functions with free knots, IMA J. Numer. Anal. 12 (1992) 95–105.
[135] W. Murray, M.L. Overton, A projected Lagrangian algorithm for nonlinear minimax optimization, SIAM J. Sci. Statist. Comput. 1 (1980) 345–370.
[136] W. Murray, M.L. Overton, A projected Lagrangian algorithm for nonlinear l_1 optimization, SIAM J. Sci. Statist. Comput. 2 (1981) 207–214.
[137] G. Nürnberger, Strongly unique spline approximation with free knots, Constr. Approx. 3 (1987) 31–42.
[138] G. Nürnberger, Approximation by univariate and bivariate splines, in: D. Bainov, V. Covachev (Eds.), Second International Colloquium on Numerical Analysis, VSP, Utrecht, 1994, pp. 143–153.
[139] G. Nürnberger, Bivariate segment approximation and free knot splines: research problems 96-4, Constr. Approx. 12 (1996) 555–558.
[140] G. Nürnberger, Optimal partitions in bivariate segment approximation, in: A. Le Méhauté, C. Rabut, L.L. Schumaker (Eds.), Surface Fitting and Multiresolution Methods, Vanderbilt University Press, Nashville, 1997, pp. 271–278.
[141] G. Nürnberger, L.L. Schumaker, M. Sommer, H. Strauss, Approximation by generalized splines, J. Math. Anal. Appl. 108 (1985) 466–494.
[142] G. Nürnberger, L.L. Schumaker, M. Sommer, H. Strauss, Uniform approximation by generalized splines with free knots, J. Approx. Theory 59 (1989) 150–169.
[143] G. Nürnberger, I. Singer, Uniqueness and strong uniqueness of best approximations by spline subspaces and other spaces, J. Math. Anal. Appl. 90 (1982) 171–184.
[144] G. Nürnberger, M. Sommer, A Remez type algorithm for spline functions, Numer. Math. 41 (1983) 117–146.
[145] G. Nürnberger, M. Sommer, H. Strauss, An algorithm for segment approximation, Numer. Math. 48 (1986) 463–477.
[146] M.R. Osborne, An algorithm for discrete, nonlinear best approximation problems, Proceedings of Oberwolfach Conference on ISNM, Vol. 16, Birkhauser, Basel, 1971.
[147] M.R. Osborne, Finite Algorithms in Optimisation and Data Analysis, Wiley, Chichester, 1985.
[148] M.R. Osborne, The reduced gradient algorithm, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amsterdam, 1987, pp. 95–107.
[149] M.R. Osborne, G.A. Watson, On the best linear Chebyshev approximation, Comput. J. 10 (1967) 172–177.
[150] M.R. Osborne, G.A. Watson, An algorithm for minimax approximation in the nonlinear case, Comput. J. 12 (1969) 63–68.
[151] M.R. Osborne, G.A. Watson, A note on singular minimax approximation problems, J. Math. Anal. Appl. 25 (1969) 692–700.
[152] M.R. Osborne, G.A. Watson, On an algorithm for discrete nonlinear L_1 approximation, Comput. J. 14 (1971) 184–188.
[153] M.R. Osborne, G.A. Watson, Nonlinear approximation problems in vector norms, in: G.A. Watson (Ed.), Proceedings of Dundee Numerical Analysis Conference, Springer, Berlin, 1978.

[154] M.R. Osborne, G.A. Watson, Aspects of M-estimation and l_1 fitting problems, in: D.F. Griffiths, G.A. Watson (Eds.), Numerical Analysis: A R Mitchell 75th Birthday Volume, World Scientific Publishing Co, Singapore, 1996, pp. 247–261.
[155] T. Pavlidis, A.P. Maika, Uniform piecewise polynomial approximation with variable joints, J. Approx. Theory 12 (1974) 61–69.
[156] G. Polya, Sur une algorithme toujours convergent pour obtenir les polynomes de meillure approximation de Tchebysheff pour une fonction continue quelconque, Comptes Rendues 157 (1913) 840–843.
[157] S. Portnoy, R. Koenker, The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators, Statist. Sci. 12 (1997) 279–296.
[158] M.J.D. Powell, Approximation Theory and Methods, Cambridge University Press, Cambridge, 1980.
[159] C.J. Price, I.D. Coope, Numerical experiments in semi-infinite programming problems, Comput. Optim. Appl. 6 (1996) 169–189.
[160] A. Ralston, Rational Chebyshev approximation by Remes algorithms, Numer. Math. 7 (1965) 322–330.
[161] A. Ralston, Some aspects of degeneracy in rational approximation, JIMA 11 (1973) 157–170.
[162] R. Reemtsen, Modifications of the first Remez algorithm, SIAM J. Numer. Anal. 27 (1990) 507–518.
[163] R. Reemtsen, Discretization methods for the solution of semi-infinite programming problems, J. Optim. Theory Appl. 71 (1991) 85–103.
[164] R. Reemtsen, S. Görner, Numerical methods for semi-infinite programming: a survey, in: R. Reemtsen, J.-J. Ruckman (Eds.), Semi-Infinite Programming, Kluwer Academic Publishers, Boston, 1998, pp. 195–275.
[165] E.YA. Remes, Sur un procédé convergent d'approximation sucessives pour déterminer les polynomes d'approximation, Comptes Rendues 198 (1934) 2063–2065.
[166] E.YA. Remes, Sur le calcul effectif des polynomes d'approximation de Tchebichef, Comptes Rendues 199 (1934) 337–340.
[167] E.C. Rhodes, Reducing observations by the method of minimum deviations, Philos. Mag. 9 (1930) 974–992.
[168] J.R. Rice, The characterization of best nonlinear Tchebycheff approximation, Trans. Amer. Math. Soc. 96 (1960) 322–340.
[169] J.R. Rice, Chebyshev approximation by $ab^x + c$, J. SIAM 10 (1960) 691–702.
[170] J.R. Rice, Tchebycheff approximation by functions unisolvent of variable degree, Trans. Amer. Math. Soc. 99 (1961) 298–302.
[171] J.R. Rice, Chebyshev approximation by exponentials, J. SIAM 10 (1962) 149–161.
[172] J.R. Rice, The Approximation of Functions, Vol. I, Addison-Wesley, Reading, MA, 1964.
[173] J.R. Rice, Characterization of Chebyshev approximation by splines, SIAM J. Math. Anal. 4 (1967) 557–567.
[174] F. Riesz, Uber lineare Funktionalgleichungen, Acta. Math. 41 (1918) 71–98.
[175] P.D. Robers, A. Ben-Israel, An interval programming algorithm for discrete linear L_1 approximation problems, J. Approx. Theory 2 (1969) 323–331.
[176] P.D. Robers, S.S. Robers, Algorithm 458: discrete linear L_1 approximation by interval linear programming, Comm. ACM 16 (1973) 629–631.
[177] H. Robitzsch, R. Schaback, Die numerische Berechnung von Startnaherungen bei der Exponentialapproximation, in: L. Collatz, G. Meinardus, H. Werner (Eds.), Numerical Methods of Approximation Theory, ISNM, Vol. 42, Birkhauser, Basel, 1978, pp. 260–280.
[178] S.A. Ruzinsky, E.T. Olsen, l_1 and l_∞ minimization via a variant of Karmarkar's algorithm, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 245–253.
[179] R. Schaback, Suboptimal exponential approximation, Report Nr 23, University of Gottingen, 1979.
[180] L.L. Schumaker, Uniform approximation by Tchebycheffian spline functions, J. Math. Mech. 18 (1968) 369–378.
[181] L.L. Schumaker, Some algorithms for the computation of interpolating and approximating spline functions, in: T.N.E. Greville (Ed.), Theory and Applications of Spline Functions, Academic Press, New York, 1969, pp. 87–102.
[182] R.R. Singleton, A method for minimizing the sum of absolute values of deviations, Ann. Math. Statist. 11 (1940) 301–310.
[183] K. Spyropoulos, E. Kiountouzis, A. Young, Discrete approximation in the L_1 norm, Comput. J. 16 (1973) 180–186.
[184] E.L. Stiefel, Uber diskrete und lineare Tschebyscheff- Approximationen, Numer. Math. 1 (1959) 1–28.
[185] E.L. Stiefel, Note on Jordan elimination, linear programming and Tchebyscheff approximation, Numer. Math. 2 (1960) 1–17.

[186] J. Stoer, A direct method for Chebyshev approximation by rational functions, J. Assoc. Comput. Mach. 1 (1964) 59–69.
[187] L. Tornheim, On n-parameter families of functions and associated convex functions, Trans. Amer. Math. Soc. 69 (1950) 457–467.
[188] K.N. Usow, On L_1 approximation I: computation for continuous functions and continuous dependence, SIAM J. Numer. Anal. 4 (1967) 70–88.
[189] K.N. Usow, On L_1 approximation II: computation for discrete functions and discretization effects, SIAM J. Numer. Anal. 4 (1967) 233–244.
[190] C.J. de la Vallée Poussin, Sur la methode de l'approximation minimum, Societe Scientifique de Bruxelles, Annales, Memoires, Vol. 35, 1911, pp. 1–16.
[191] L. Veidinger, On the numerical determination of the best approximations in the Chebyshev sense, Num. Math. 2 (1960) 99–105.
[192] H.M. Wagner, Linear programming techniques for regression analysis, J. Amer. Statist. Assoc. 54 (1959) 206–212.
[193] J.L. Walsh, The existence of rational functions of best approximation, Trans. Amer. Math. Soc. 33 (1931) 668–689.
[194] G.A. Watson, A multiple exchange algorithm for multivariate Chebyshev approximation, SIAM J. Numer. Anal. 12 (1975) 46–52.
[195] G.A. Watson, A method for calculating best nonlinear Chebyshev approximations, JIMA 18 (1976) 351–360.
[196] G.A. Watson, On two methods for discrete L_p approximation, Computing 18 (1977) 263–266.
[197] G.A. Watson, On a class of methods for nonlinear approximation problems, in: D.C. Handscomb (Ed.), Multivariate Approximation, Academic Press, London, 1978, pp. 219–227.
[198] G.A. Watson, The minimax solution of an overdetermined system of nonlinear equations, JIMA 23 (1979) 167–180.
[199] G.A. Watson, Approximation Theory and Numerical Methods, Wiley, Chichester, 1980.
[200] G.A. Watson, An algorithm for linear L_1 approximation of continuous functions, IMA J. Numer. Anal. 1 (1981) 157–167.
[201] F. Wenzl, Uber Gleichungssysteme der Tschebysheffschen Approximation, Z. Angew. Math. Mech. 34 (1954) 385–391.
[202] H. Werner, Die konstruktive Errmittlung der Tschebyscheff-Approximation in Bereich der rationalen Funktionen, Arch. Rational Mech. Anal. 11 (1962) 368–384.
[203] H. Werner, Rationale-Tschebyscheff-Approximation, Eigenwerttheorie und Differenzenrechnung, Arch. Rational Mech. Anal. 13 (1963) 330–347.
[204] H. Werner, Der Existenzsatz fur das Tschebyscheffsdhe Approximationsproblem mit Exponenttialsummen, in: L. Collatz, H. Unger (Eds.), Funktionalanalytische Methoden der numerischen Mathematik, ISNM, Vol. 12, Birkhauser, Basel, 1969, pp. 133–143.
[205] H. Werner, Tschebysheff-Approximation with sums of exponentials, in: A. Talbot (Ed.), Approximation Theory, Academic Press, London, 1969, pp. 109–134.
[206] J.M. Wolfe, On the convergence of an algorithm for discrete L_p approximation, Numer. Math. 32 (1979) 439–459.
[207] R.S. Womersley, R. Fletcher, An algorithm for composite nonsmooth optimization problems, J. Optim. Theory Appl. 48 (1986) 493–523.
[208] J.W. Young, General theory of approximation by functions involving a given number of arbitrary parameters, Trans. Amer. Math. Soc. 8 (1907) 331–344.
[209] Y. Zhang, A primal-dual interior point approach for computing the l_1 and l_∞ solutions of overdetermined linear systems, J. Optim. Theory and Appl. 77 (1993) 323–341.
[210] J.J. Zhou, A.L. Tits, An SQP algorithm for finitely discretized continuous minimax optimization problems and other minimax problems with many objective functions, SIAM J. Optim. 6 (1996) 461–487.
[211] S.I. Zuhovickii, L.I. Avdeyeva, Linear and Convex Programming, (Translated from the 1964 Russian edition), Saunders Co., Philadelphia, 1966.
[212] S.I. Zuhovickii, R.A. Poljak, M.E. Primak, An algorithm for the solution of the problem of Cebysev approximation, Soviet. Math. Dokl. 4 (1963) 901–904.