The University of Southwestern Louisiana

Lafayette, Louisiana

Proceedings of the

Thirteenth Annual

USL
Mathematics Conference

October 22-24 1982

Edited by
Dr. Baker Kearfott

May, 1984

PROCEEDINGS OF THE THIRTEENTH ANNUAL

USL MATHEMATICS CONFERENCE

October 22-24, 1982

Organized by

Baker Kearfott

Grace Poeling

and

Lloyd Poeling

Edited by

Ralph Baker Kearfott

Department of Mathematics

University of Southwestern Louisiana

July, 1983

LIST OF SPEAKERS

Yalcin Acar (Louisiana State University):
"Interpetation of the Dissipation of Penetrtion Pore Pressures"

Ward Cheney (University of Texas at Austin):
"Approximation in Tensor Product Spaces"

John Dennis (Rice University):
"Secant Type Methods for Noisy Functions"

Andreas Griewank (Southern Methodist University)

David Kincaid (University of Texas at Austin):
"Supercomputers and Solving Linear Systems Using Iterative Methods"

Beny Neta (Texas Tech University):
"Higher-Order Methods for Determining Nonisolatee Solutions of Nonlinear Equations"

Larry L. Schumaker (Texas A & M University):
"Spaces of Piecewise Polynomials in Two Variablet"

Vijay Singh (Louisiana State University):
"Mixed Solutions to Hydraulic Equations for Flow over a Porous Bed"

Gary Sod (Tulane University)

Arthur Stroud (Texas A & M University):
"Interactive Quadrature"

Mary Wheeler (Rice University):
"Petroleum Reservoir Modeling"

David Young (University of Texas at Austin):
"On a Simplification of Generalized Conjugate Gradient Methods for Nonsymmetric Linear Systems"

CONTENTS

INTERPRETATON OF THE DISSIPATION OF PENETRATION PORE PRESSURES

Yalcin B. Acar
Louisiana State University, Baton Rouge, Louisiana, USA

Mehmet T. Tumay
Louisiana State University, Baton Rouge, Louisiana, USA

Adrian Chan
Louisiana State University, Baton Rouge, Louisiana, USA

1 INTRODUCTION

A wide variety of insitu testing methods exists for determining the average time deformation characteristics of soils. Among others, constant head or falling head permeameter tests are the most favored (Milligan 1975). However, the initial waiting period needed for dissipation of pore pressure generated during driving and installation of the equipment together with disturbance of the deposit make these methods expensive and sometimes misleading. The piezometer probe introduced by Wissa, et al. (1975) and Torstensson (1975) continuously records the pore pressures generated during penetration. Due to their smaller dimensions and sharp edges, these probes generate less disturbance that the permeameters. The dissipation of penetration pore pressures after stopping the probe is related to the consolidation characteristics of the deposit.

Piezometric elements have recently been incorporated in the cone penetrometer, thus permiting the simultaneous recording of generated pore pressures, u. Continuous pore pressure records from piezocone soundings have extensively contributed to the understanding of the variation of tip resistance, q_c, and sleeve friction, f_s, values with soil type and properties (Tumay, et al. 1981). The analysis of the dissipation of these pore pressures permit a new and economical method of estimation of the coefficient of consolidation and coefficient of permeability. An accurate analysis of the field dissipation phenomenon around cones require a coupled, nonlinear and axisymmetric analysis of consolidation. Such an analysis necessitates a thorough knowledge of the initial conditions of the problem which consists of the spatial distribution of pore pressures and total stresses before dissipation together with the nonlinear behavior of the deposit. However, uncertainties involved in estimation of these initial conditions and the different variables that affect them make such an analysis unduely complex.

This study presents a method of predicting the insitu coefficient of consolidation from the pore pressure dissipations when piezo-cone penetration is stopped. In view of the experience attained from insitu dissipation studies in very soft cohesive deposits in Louisiana, different factors that affect dissipation are studied with a parametric FEM analysis of uncoupled, linear consolidation around cones.

2 PENETRATION PORE PRESSURES

Pore pressures generated during penetration constitute the only initial conditions for a linear uncoupled analysis of the dissipation around penetrometers. Furthermore, in an uncoupled, linear analysis the magnitude of these pore pressures is ignored since consolidation is governed by the axisymmetric diffusion equation,

$$\nabla^2 U = \frac{\partial U}{\partial T} \qquad (1)$$

where U and T are normalized dimensionless variables defined as,

$$U = u/u_1 \qquad (2)$$

$$T = \bar{c}t/R^2 \qquad (3)$$

$$\bar{c} = Ek/\gamma_w \qquad (4)$$

where u = pore pressure at a given time, u_i = initial pore pressure, $\bar{c}$ = coefficient of consolidation, t = time, R = radius of cone, E = deformation modulus, k = hydraulic conductivity, and γ_w = unit weight of water.

It is noted that the dissipation is governed by the initial, spatial distribution of pore pressures surrounding the cone. Variation of penetration pore pressures along the shaft of the penetrometer is obtained by emplacing the piezometer element at different locations on the cone. Figure 1 presents the experience obtained by different investigators. These results indicate that the tip values of pore pressures are approximately 2.5 times the shaft values. Although it is possible to record these pore pressure at different locations on the shaft and tip of the cone, efforts to obtain the spatial distribution have rendered erroneous results due to changes in the boundary conditions of the problem by the insertion of a second probe (Baligh and Levadoux 1980). However, experience in the field has pointed out the fact that in very soft cohesive soils, penetration pore pressures extend to 20-25 times the radius of the probe. Similar results were obtained in measuring the pore pressure generated during pile driving (Roy, et al. 1979). Levadoux and Baligh (1980) approached the

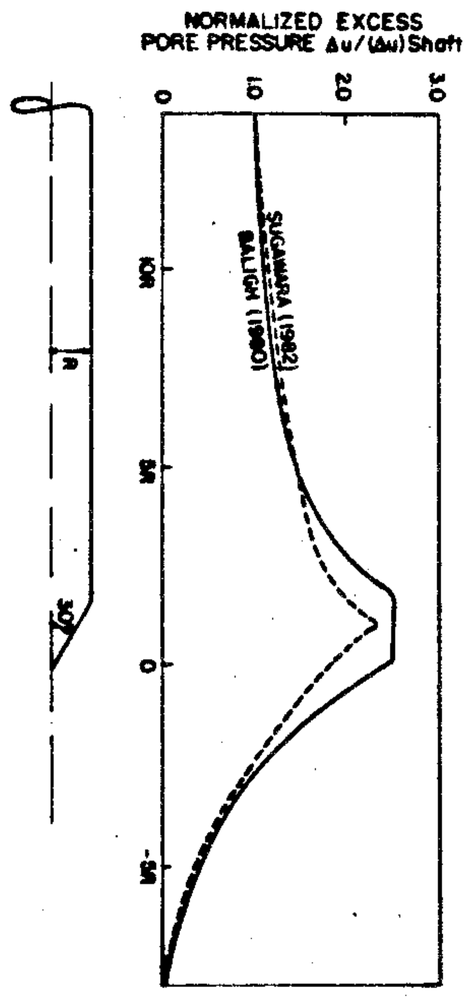

Figure 1. Variation of penetration pore pressures along the shaft.

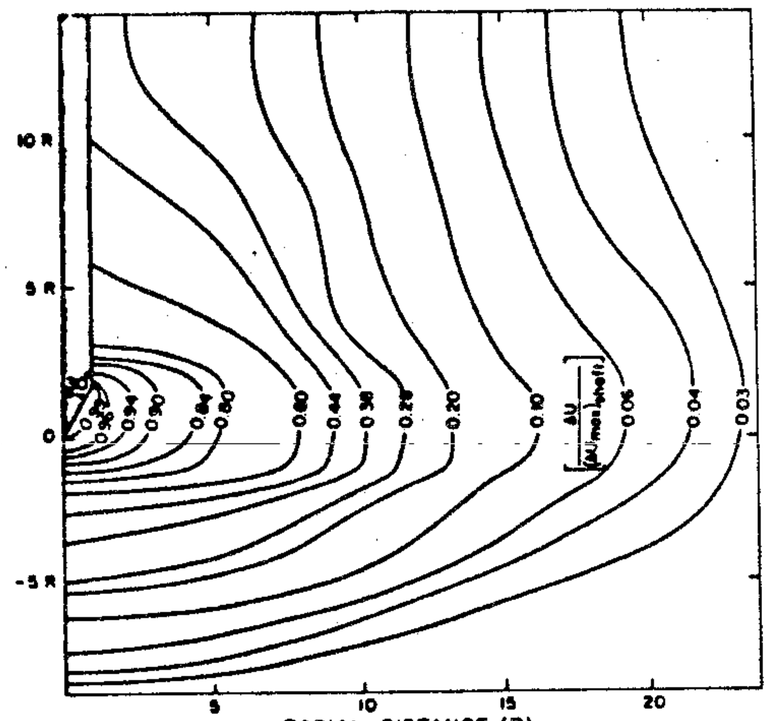

Figure 2. Spatial distribution of normalized pore pressures during penetration.

problem by estimating the strain rates induced by a cone penetrating an inviscid and incompressible fluid and approximating the pore pressures generated from this strain field by the soil plasticity models. Theoretical approximations and field measurements indicate that the spatial distribution of pore pressures along the shaft can best be estimated from the results of data obtained from pile drivings (Chan 1982).

Figure 2 presents the spatial distribution of normalized pore pressures during penetration. It is observed that the pore pressures at the tip form a plateau for a distance of about 4 R after which they decrease logarithmically with radial distance. An axisymmetric FEM code was written in order to estimate the dissipation of these initial pore pressures. It was determined that taking the free draining boundary at a distance of about 25 R closely approximates an infinite radial boundary. Therefore, the mesh given in Figure 3 was utilized. The top and bottom of the mesh are taken as impervious since the flux out of the boundaries at these distances is negligible.

3 ANALYSIS OF DISSIPATION

It is firstly attempted to analyze the effect of anisotropy of a soil deposit on the pore pressure dissipation data. Figure 4 shows the effect of anisotropy

in hydraulic conductivity. The time factor is given in horizontal coefficient of consolidation in this figure. It is observed that the consolidation around a cone is basically governed by the horizontal time deformation characteristics of the deposit. Furthermore, the effect of anisotropy is least pronounced during the final stages of consolidation. This suggests that the insitu time readings taken from the final stages of dissipation would correlate better with the horizontal value of coefficient of consolidation. However, waiting for full dissipation of initial excess pore pressure in practice might interrupt the soundings anywhere from one to thirty hours. This requirement of waiting for later stages of dissipation may be quite impractical and uneconomical. To remedy this shortcoming of the test and still obtain a meaningful value of dissipation time, it is recommended that an approximate curve fitting technique be used. It is shown in Figure 4 that when the straight portion of the dissipation curves is extended to 90 percent dissipation, the ratio of the actual time factor, T_{90}, to the extrapolated T'_{90}, could be taken as a constant,

$$(\log T_{90})/(\log T'_{90}) = A \qquad (5)$$

where $T_{90} = 240$, $T'_{90} = 130$ and $A = 1.13$ is obtained. Consequently, in a field dissipation curve, the actual time of 90 percent dissipation could be estimated by,

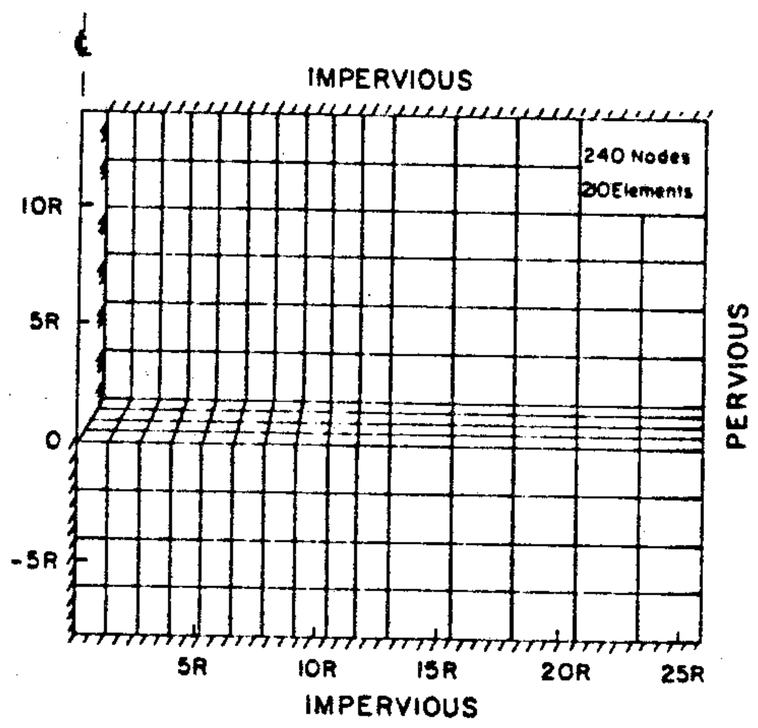

Figure 3. Finite element mesh for axisymmetric dissipation around penetrometer.

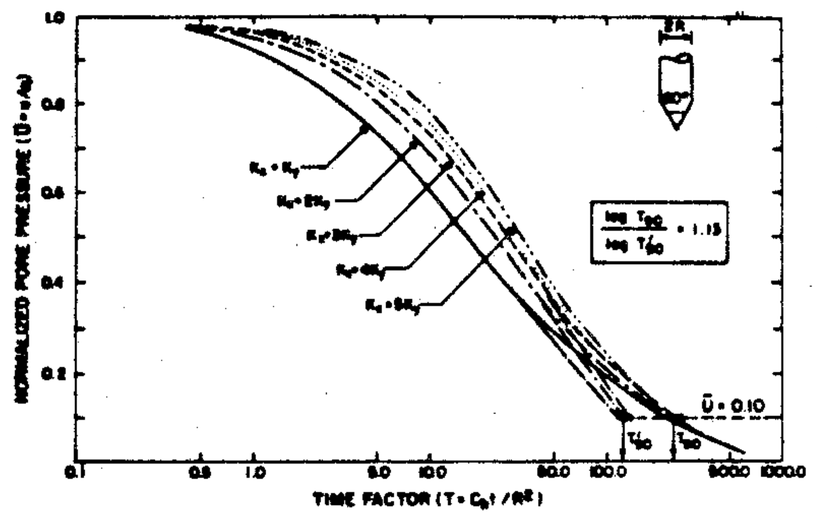

Figure 4. Effect of anisotropy on dissipation and estimation of 90 percent dissipation time.

$$t_{90} = 10^{1.13 \log t'_{90}} \qquad (6)$$

where t'_{90} = the time to 90 percent dissipation when the dissipation is extrapolated as a straight line and t_{90} = actual time of 90 percent dissipation.

Sharper cones are sometimes used since they impose less disturbance to soil and generate lower magnitudes of initial pore pressures. Figure 5 gives the dissipation results for an 18° and 60° cone. It is observed that at intermediate stages of dissipation longer time is required for sharper cones while the two curves converge for final stages of dissipation. This implies that the dissipation tests with sharper cones will take a longer time at initial stages of consolidation.

Cone penetration imposes very high straining to the soil in front of the tip [50 percent strain level as predicted by Baligh and Levadoux (1980)]. This suggests that when the penetration is stopped, there exists a highly disturbed inner zone and a rather undisturbed outer zone surrounding the cone. High strains

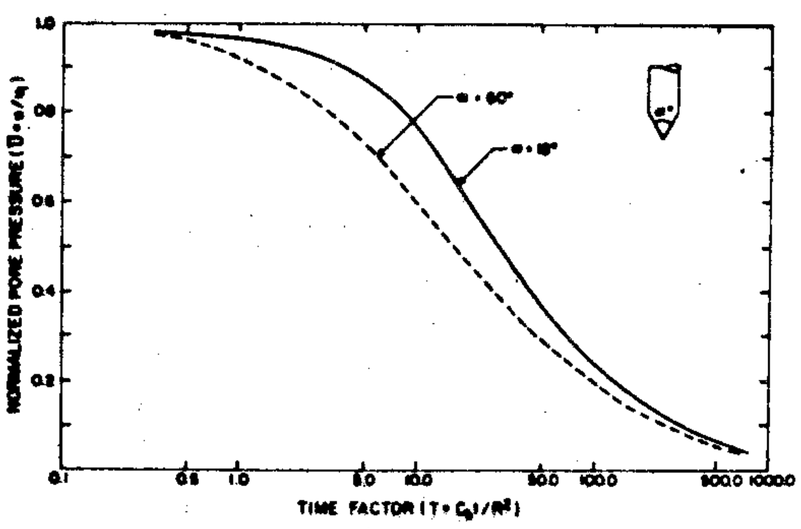

Figure 5. Effect of cone angle on dissipation.

around the tip extend to regions at a distance of about 5 R after which the strain level could be approximated by cavity expansion theories and which is in the order of 1-2 percent. Therefore, it is logical to assume that the coefficient of consolidation of the region surrounding the tip will be higher than the relatively less disturbed outer zone. The dissipation results for different ratios of the coefficient of consolidation in these two zones are presented in Figure 6. It is observed that mainly the properties of the undisturbed region dominate the dissipation while at final stages of consolidation all dissipation results converge.

Finally the effect of recording the dissipation at different positions on the shaft is investigated. Figure 7 shows that it is advantageous to put the piezometric element at the tip rather than the shaft. This would save a considerable amount of time during the dissipation tests at the field. Furthermore, it is noted that the dissipation of normalized pore pressures for piezometric elements at any location on the tip are almost identical and consequently results obtained from cones with piezometric elements at the tip or at the middle of the tip should be comparable in analysis of dissipation records.

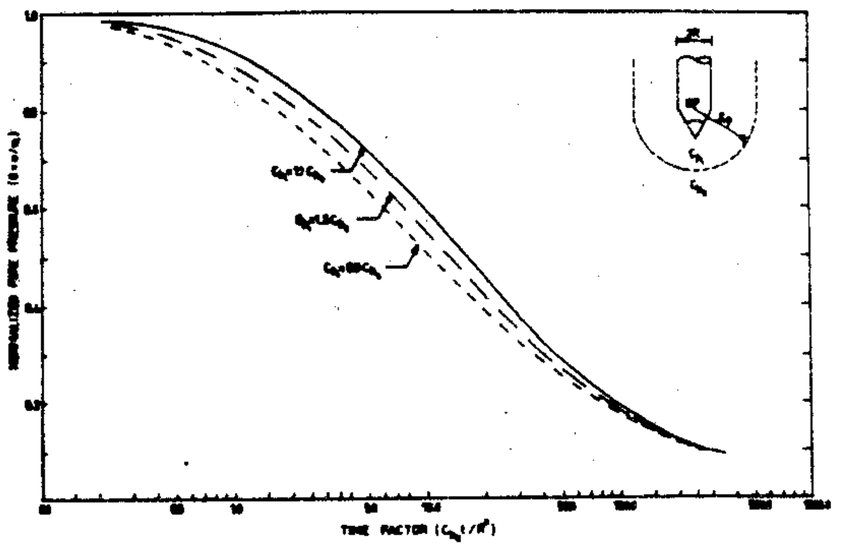

Figure 6. Effect of disturbed zone on dissipation.

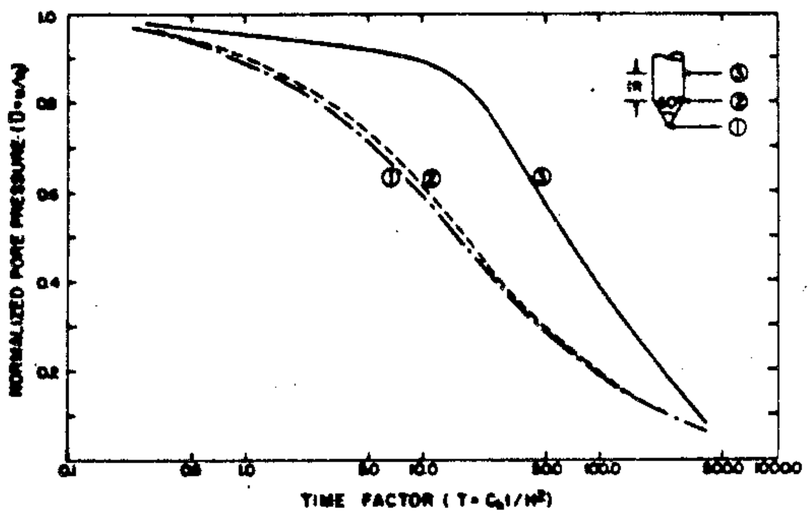

Figure 7. Dissipation at different locations on cone penetrometers.

4 EVALUATION OF RESULTS

Figure 8 presents the tip resistance and pore pressure records in an alluvial deposit of southern Louisiana. It is observed that the upper highly compressible clay deposit is separated by a rather silty clay layer from a stiffer lower clay deposit. Three dissipation records were taken in both the upper and the lower deposits. The dissipation of total pore pressures with time is given in Figure 9. These total pore pressures consist of both the excess pore pressures, Δu_i, and the hydrostatic pore pressures, u_o. The ratio of the excess pore pressure to the tip resistance, $\Delta u_i/q_c$, is a good indication of the stress history of the deposit (Tumay, et al. 1981). In conformity with the boring logs, it is noted that higher values of $\Delta u_i/q_c$ obtained in the upper deposit indicate lower overconsolidation ratios.

The parallel dissipation curves in Figure 9 indicate that the spatial distribution of penetration pore pressures were similar while the average coefficient of consolidation values are

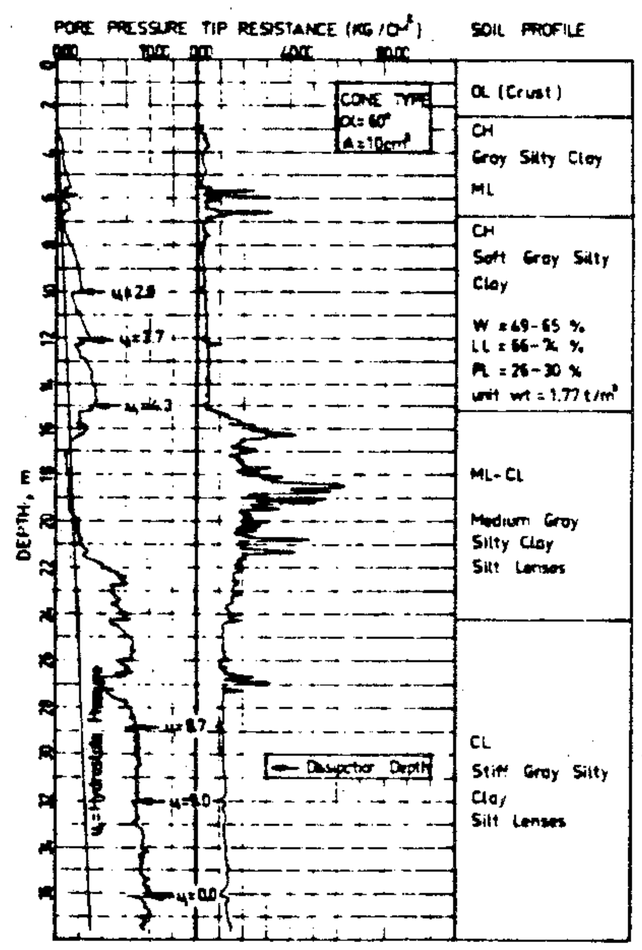

Figure 8. Penetration records and boring log in Norco, Louisiana.

different. The anomaly presented by dissipation curve 6 might be due to the presence of a silt lens close to the penetration location. Due to this possible inhomogeneity around the cone, the initial normalized pore pressure distribution at this location might be different than the distribution taken in numerical analysis. Consequently, this dissipation will not be evaluated for the coefficient of consolidation.

The excess pore pressure in the first five dissipation records which are normalized with respect to the initial maximum pore pressures, are shown in Figure 10. With regard to the discussion presented earlier, the straight portion of the dissipation records in the logarithmic plot is extended down to 90 percent dissipation and the corresponding actual dissipation time is calculated by equation 6. The time factor for 90 percent dissipation is taken as 240 and the c_h values are calculated. The values presented in Figure 10 are found to be generally 10 to 20 times the values obtained from laboratory results for vertical coefficient of consolidation (Chan 1982). However, it should be noted that during the different stages of dissipation, some regions of the soil surrounding the cone is subjected to both a swelling and recompression. This behavior might contribute to the higher average c_h values obtained. Furthermore, Al-Dhahir, et al. (1970) observed that the coefficient of consolidation values backcalculated from the insitu evaluation of pore pressures under an embankment in very soft cohesive soils, are generally close to values obtained from insitu permeameter tests while laboratory consolidation data have given values 10-20 times smaller. In the light of these studies, it could be concluded that dissipation results obtained by piezo-cone tests give a more reliable estimation of the insitu horizontal coefficient of consolidation.

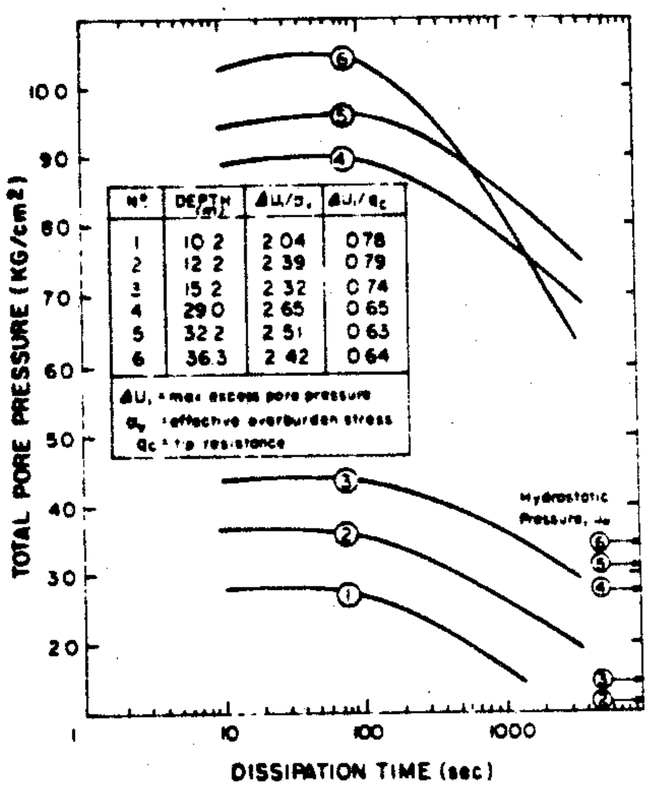

Figure 9. Field dissipation records.

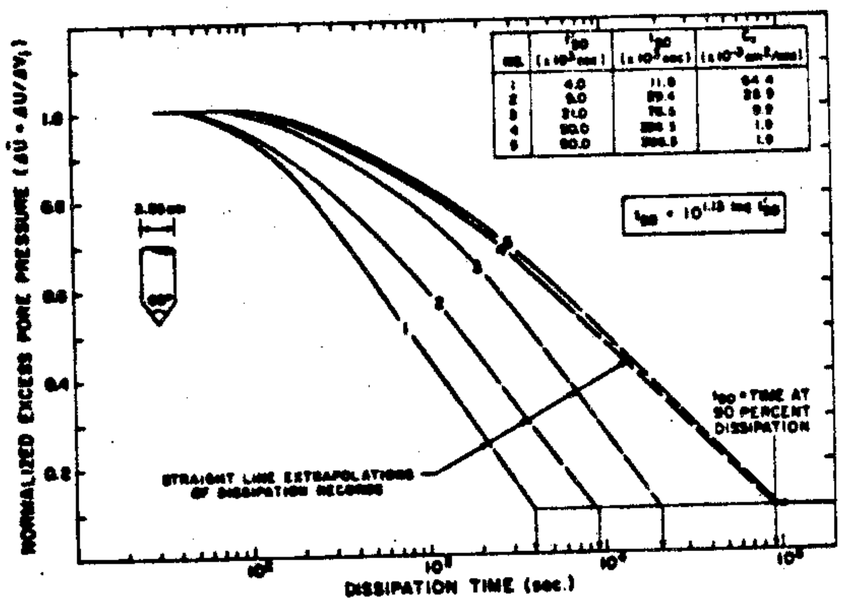

Figure 10. Normalized field dissipation records.

5 SUMMARY AND CONCLUSIONS

A method to interpret the dissipation of penetration pore pressures in piezo-cone penetration testing is presented. It is noted that linear, uncoupled FEM analysis of the dissipation of penetration pore pressures yield reliable values of coefficient of consolidation for soft cohesive soils. Analysis of numerical studies indicate that:

1. The dissipation of pore pressure due to cone penetration is governed mainly by the horizontal values of coefficient of consolidation. It is best to evaluate the results of this test at later stages in order to reduce the influence of anisotropy due to hydraulic conductivity.
2. Sharper cones require longer dissipation times, hence a 60° cone angle is recommended for routine dissipation tests since they would take less dissipation time.
3. Although the soil immediately surrounding the cone is highly disturbed, dissipation results at the later stages of consolidation reflect the characteristics of the undisturbed deposit.
4. Faster dissipation curves are obtained when the pore pressures are

recorded at any position on the conic part of the penetrometer. Therefore it is preferrable that the piezometric elements be placed at these positions for dissipation studies.

5. Normalized dissipation curves will be similar for piezometric elements emplaced at the tip or at the middle of the tip. This suggests that the dissipation results obtained from cones with the same apex angle but with piezometric elements at different positions of the tip must be comparable.

6. Coefficient of consolidation obtained by piezo-cone penetration tests would generally be higher than values obtained from conventional oedometer tests. This is mainly attributed to both the uncertainties involved in the initial distribution of pore pressures during the test and the general shortcomings of performing a laboratory consolidation test. However, previous studies on insitu evaluation of volume change characteristics of soft cohesive soils indicate that coefficient of consolidation obtained by insitu tests are close to actual values but generally greater than laboratory values (Mitchell & Gardner 1975). This indicates that the dissipation tests give better estimations of the coefficient of consolidation in a homogeneous deposit.

6 ACKNOWLEDGEMENTS

Dr. Roger K. Seals is gratefully acknowledged for his helpful suggestions and critical review of this manuscript.

The Louisiana Department of Transportation and Development (LDODT) is acknowledged for the financial support provided for this study. The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the LDOTD or the Federal Highway Administration. This paper does not constitute a standard, specification or regulation.

REFERENCES

Al-Dhahir, Z.A., M.F.Kennard & N.R. Morgenstern 1970. Observations on pore pressures beneath the ash lagoon embankments at Fiddler's Ferry Power Station. Proceedings, Conf. on In-Situ Investigations in Soils and Rocks, British Geotechnical Soc., London, Paper No. 20, 265-276.

Baligh, M.M. & J.N.Levadoux 1980. Pore pressure after cone penetration. Research Report, MITSG 80-13, Cambridge, MA.

Chan, A. 1982. Analysis of dissipation of pore pressures after cone penetration. MS Thesis, Dept. of Civil Engr., Louisiana State University.

Levadoux, J.N., M.M.Baligh 1980. Pore pressure during cone penetration in clays. Research Report, Dept. of Civil Engr., MIT, Cambridge, MA.

Milligan, V. 1975. Field measurement of permeability in soil and rock. Proceedings, ASCE Specialty Conf. on In-Situ Measurements of Soil Properties, Raleigh, NC, 1:3-36.

Mitchell, J.K. & W.S.Gardner 1975. In situ measurement of volume change characteristics. Proceedings, ASCE Specialty Conf. on In Situ Measurement of Soil Properties, Raleigh, NC, 1:279-345.

Roy, M., et al. 1979. Behavior of a sensitive clay during pile driving. Proceedings, Canadian Geotechnical Conf., Quebec, 4:28-49.

Sugawara, N. & M.Chikaraishi 1982. On Estimation of ø' for normally consolidated mine tailings using the pore pressure cone penetrometers. Proceedings, Second European Symposium on Penetration Testing (ESOPT II), Amsterdam, The Netherlands, 2:883-888.

Torstensson, B.A. 1975. Pore pressure sounding instrument. Proceedings, ASCE Specialty Conf. on In-Situ Measurements of Soil Properties, Raleigh, NC, 2:48-54.

Tumay, M.T., R.L.Boggess & Y.Acar 1981. Subsurface investigation with piezo-cone penetrometer. Proceedings, ASCE Session on Cone Penetration Testing and Experience, St. Louis, MO, 325-342.

Wissa, A.E.Z., R.T.Martin & J.E.Garlenger 1975. The piezometer probe. Proceedings, ASCE Specialty Conf. on In-Situ Measurements of Soil Properties, Raleigh, NC, 1:536-545.

Ενε

THE BEST APPROXIMATION OF MULTIVARIATE FUNCTIONS BY COMBINATIONS OF UNIVARIATE ONES

E. W. Cheney

This survey concerns various schemata for approximating multivariate functions by combinations of univariate functions. The combinations of interest are formed by addition, multiplication, functional composition, and passage to a limit. Some very intractable problems are outlined, but our emphasis is on problems in which some progress has been made in recent years. These are mainly the cases when the approximating functions form linear subspaces. In contrast with the situation in classical approximation theory, the subspaces which occur in this subject are generally of infinite dimension.

1. Introduction

The theorem of Arnold and Kolmogoroff informs us that any continuous multivariate function can be constructed by using continuous univariate functions as "building blocks", and the operations of composition and addition as "mortar". The Weierstrass Approximation Theorem provides similar information; the "building blocks" are the same, and the "mortar" consists of multiplication, addition, and passage to the limit. Here are the 2-variable forms of these representations:

$$f(x,y) = \sum_{i=1}^{5} \phi_i(g_i(x) + h_i(y)) \tag{1}$$

$$f(x,y) = \sum_{i=1}^{\infty} g_i(x)h_i(y) \ . \tag{2}$$

In both of these equations, further stipulations may be made. For example, we can take all ϕ_i to be the same, and h_i to be a multiple of g_i in (1). In (2), we can assume that the functions g_i and h_i are polynomials. References [32] and [44] are convenient ones for the theorem of Kolmogorov and Arnold.

In general, if we restrict the flexibility of the functional forms on the right side of (1) and (2), the possibility of <u>exact representation</u> will be lost, and what remains is an interesting problem of <u>best</u> approximation. Thus, for example, we can seek a best approximation to a function f by functions of the form

$$\sum_{i=1}^{4} \phi_i(g_i(x)+h_i(y)) \quad \text{or} \quad \sum_{i=1}^{n} g_i(x)h_i(y) \ .$$

These two problems, arrived at in an innocent attempt to replace exact representations by approximations, are in fact quite formidable, and no one has suggested methods for attacking them numerically.

One point that should be emphasized is that when we give up exact representations and search for "best approximations" the norm which is used will have a profound influence on the problem. As long as our approximation problem is linear (and in some rare instances of nonlinear approximation) the L_2-norm leads to the simplest theory. More interesting problems occur with the L_∞ and L_1 norms.

The remainder of this survey will be organized around certain specific forms of approximation. These can be enumerated as follows. In each case, f is a function to be approximated by a function of the type appearing on the right side of the $\approx$ sign.

(A) $f(x,y) \approx g(x)$

(B) $f(x,y) \approx g(x) + h(y)$

(C) $f(x,y) \approx \phi(g(x)+h(y))$

(D) $f(x,y) \approx \sum_{i=1}^{n} g_i(x)h_i(y)$

(E) $f(x,y) \approx \sum_{i=1}^{n} u_i(x)h_i(y) + \sum_{i=1}^{m} v_i(y)g_i(x)$

Each of these problems will have a number of variants, which we will label A1, A2, and so forth.

2. Problem (A)

We begin with Problem (A1), in which a continuous function f on $X \times Y$ is given. Here X and Y can be arbitrary compact Hausdorff spaces. We seek an element $g \in C(X)$ to minimize the expression

$$\|f-g\| = \sup_y \sup_x |f(x,y)-g(x)| \ . \tag{3}$$

Let us introduce the concept of the sections of a bivariate function. These are the univariate functions which one obtains by fixing one argument. The standard notation is adopted, viz.

(4) $$f_x(y) = f^y(x) = f(x,y) \ .$$

It is an exercise in equicontinuity to prove that if $f \in C(X \times Y)$, then the set $K = \{f^y : y \in Y\}$ is a compact subset of $C(X)$; i.e., it is bounded, closed, and equicontinuous. These considerations allow us to re-write Equation (3) as,

(5) $$\|f-g\| = \sup_y \|f^y - g\| = \sup_{k \in K} \|k-g\| \ .$$

The element $g \in C(X)$ which makes this a minimum is called the Tchebycheff center of the set K. We would also term g the best simultaneous approximation of K. Other authors have used the term "global approximation".

The Tchebycheff center of any (bounded) set A in any Banach space F can be defined as

(6) $$E(A) = \{g \in F : \sup_{a \in A} \|a-g\| = r(A)\}$$

where $r(A)$ is the Tchebycheff radius, given by

(7) $$r(A) = \inf_{g \in F} \sup_{a \in A} \|a-g\| \ .$$

THEOREM 1. Let $C(X)$ be the Banach space of all bounded continuous real-valued functions on an arbitrary topological space X. The Tchebycheff center of any bounded subset of $C(X)$ is nonempty.

This theorem was proved for compact spaces by Zamjatin [56] and extended by Franchetti and Cheney [17].

Some important papers devoted to this topic are [1],[21],[37],[54], and [36].

The problem of minimizing the expression in Equation (4) by appropriate choice of g in $C(X)$ is easier than the general problem of finding a Tchebycheff center because the set K is compact. In fact, one solution is obtained by defining

(8) $$g(x) = \frac{1}{2} \max_{y \in Y} f(x,y) + \frac{1}{2} \min_{y \in Y} f(x,y) \ .$$

A more difficult problem of the same type we label as Problem (A2). Here g is not allowed to range over all of $C(X)$ but is restricted to some prescribed subspace, G. This leads to the concepts of restricted

Tchebycheff radius and restricted Tchebycheff center. These are defined as follows.

$$r_G(A) = \inf_{g\in G} \sup_{a\in A} \|a-g\| \tag{9}$$

$$E_G(A) = \{g \in G: \sup_{a\in A} \|a-g\| = r_G(A)\} \ . \tag{10}$$

Here A can be a bounded subset in any Banach space F, and G can be any subspace of F. A duality theorem, proved in [17], goes as follows:

THEOREM 2. *If A is compact then*

$$r_G(A) = \max_{\phi\in G^{\perp}} \inf_{\phi(f)=0} \sup_{a\in A} \|a-f\| \ . \tag{11}$$

Smith and Ward proved in [54] this important result:

THEOREM 3. *If A is a bounded set in $C(X)$, with X an arbitrary topological space, and if G is any subset of $C(X)$, then*

$$r_G(A) = r(A) + \operatorname{dist}(G,E(A)) \ . \tag{12}$$

A subset G in a Banach space F is said to be "proximinal" if each element of F has at least one best approximation in G. If we assume in Theorem 3 that G is proximinal, then a necessary and sufficient condition for the Tchebycheff center $E_G(A)$ to be nonempty is that the expression $\operatorname{dist}(f,G)$ should attain its infimum as f ranges over $E(A)$. The interested reader should consult [54] and [17] for further results in this area. One problem which remains open is this one:

CONJECTURE. *If G is a proximinal set in $C(X)$, then it is also proximinal in $C(X\times Y)$ for arbitrary compact Y.*

If this conjecture is correct, then the infimum of $\|f-g\|$ will be attained as g ranges over G, for each $f \in C(X\times Y)$. The conjecture is false if Y is allowed to be a noncompact space.

3. Problem (B)

Here we fix an $f \in C(X \times Y)$, with X and Y compact, and we seek $g \in C(X)$ and $h \in C(Y)$ to make the expression

$$\|f-g-h\| = \sup_{x,y} |f(x,y)-g(x)-h(y)| \tag{13}$$

an absolute minimum. This we term Problem (B1). Much is known about it. Diliberto and Straus were the first to study it [13], and later Aumann [3] and Golomb [27] contributed to it. For recent work and more bibliographic detail see [24] and [42].

To summarize the important known results, we state the following theorem.

THEOREM 4. The above problem always has a solution (g,h). One solution can be obtained by the Diliberto-Straus Algorithm described below.

The algorithm is an iterative one given by the formulas $f_1 = f$, $f_{n+1} = f_n - P_n f_n$ where P_n is a simple averaging operator previously encountered in Equation (8). Specifically, for even n,

$$(P_n f)(x,y) = \frac{1}{2} \max_s f(x,s) + \frac{1}{2} \min_s f(x,s) \tag{14}$$

while for odd n,

$$(P_n f)(x,y) = \frac{1}{2} \max_s f(s,y) + \frac{1}{2} \min_s f(s,y) \ . \tag{15}$$

Thus in either case $P_n f$ is really a univariate function. It is known that the sequence f_n converges uniformly, and $f - \lim f_n$ is a best approximation to f of the form $g(x)+h(y)$.

This algorithm is identical to one known as the von Neumann Alternating Method. It was described by von Neumann in 1933 and was used for obtaining best approximations in Hilbert space from the vector sum of two subspaces. See [48] for this application. A recent paper by Deutsch contains further results on this algorithm as well as further references [11].

Problem (B1) has applications to the scaling of matrices (for the purpose of preconditioning them). See [4],[23].

It is possible to characterize solutions to Problem (B1) in the following way:

THEOREM 5. In order that the pair (g,h) solve the minimum problem in Equation (13) it is necessary and sufficient that there exist an "alternant".

An "alternant" is an infinite sequence of points $p_n = (x_n, y_n)$ such that $x_n = x_{n+1}$ (and $y_n \neq y_{n+1}$) when n is odd, $y_n = y_{n+1}$ (and $x_n \neq x_{n+1}$) when n is even, and $(f-g-h)(p_n) = (-1)^n \|f-g-h\|$. This theorem is due to Havinson [31] and has been recently generalized to encompass Problem (E).

Von Neumann's algorithm is designed to produce best approximations in Hilbert space. One is presented with two closed linear subspaces U and V in Hilbert space, and has at hand the orthogonal projections onto U and V. Then the sequence $f_{n+1} = f_n - P_n f_n$ $(n = 0, 1, 2, \ldots)$ is generated, where P_{2k} is the orthogonal projection on U and P_{2k+1} is the orthogonal projection on V. Then $f_0 - \lim f_n$ is the best approximation of f_0 in the closure of $U+V$.

It is known that the algorithm works also in any uniformly convex Banach space whose dual is also uniformly convex; however, one must assume as hypothesis that $U+V$ is closed. It seems to be an open problem whether this hypothesis is essential. This result has been given by Deutsch [11]. See also [19] and [20].

An interesting question about the space $C(X \times Y)$ is this: what is the (relative) projection constant of the subspace $C(X)+C(Y)$, and what are the minimal projections? This question has recently been answered by Jameson and Pinkus, [34]. Their result is

THEOREM 6. *If X and Y are compact Hausdorff spaces, each containing infinitely many points, then the projection constant of $C(X)+C(Y)$ as a subspace of $C(X \times Y)$ is* 3.

A minimal projection, P, in this situation is obtained by selecting $\xi \in X$, $\eta \in Y$, and defining

$$Pf = f^{\eta} + f_{\xi} - f(\xi, \eta) \ . \tag{16}$$

An important variant of Problem (B), designated as (B2), arises when f is defined not on $X \times Y$ but on a compact subset S in $X \times Y$. This complicates the problem considerably, and a paper of Ofman [49] is devoted to the question of existence of a best approximation. His principal theorem is this:

THEOREM 7. *Let S be a closed subset of $X \times Y$ containing a point (ξ, η) such that (x, η) and (ξ, y) belong to S whenever (x, y) belongs to S. Then each f in $C(S)$ which satisfies a Lipschitz condition has a best approximation in the class $\{g+h: g \in C(X), h \in C(Y)\}$.*

Another important variant of Problem (B), designated (B3), involves the L_1-norm. The most natural question to ask is whether each $f \in L_1(X \times Y)$ necessarily has a best approximation (in the L_1-norm) of the form $g(x) + h(y)$, with $g \in L_1(X)$ and $h \in L_1(Y)$. The following result has been proved in [33].

THEOREM 8. Let X and Y be measure spaces of finite measure, and let $f \in L_1(X \times Y)$. Then f possesses a best L_1-approximation in the subspace $M = \{g+h: g \in L_1(X), h \in L_1(Y)\}$.

The appropriate variant of the Diliberto-Straus algorithm will not always work. Various hypotheses (either on the function being approximated or on the measure spaces X and Y) can be introduced to obtain a convergence theorem. For example, W.A. Light has proved the following result [39]:

THEOREM 9. In addition to the hypotheses of Theorem 8, assume that f differs almost everywhere from each function in M. Then the iterates f_n produced by the L_1-version of the Diliberto-Straus algorithm satisfy $\|f_n\| \downarrow \operatorname{dist}(f,M)$.

4. Problem (C)

In Problem (C1), we seek three univariate functions, φ, g, h, so that $f(x,y)$ is well-approximated by $\phi(g(x)+h(y))$. A function of the latter form (composed of three continuous functions) is said to be nomographic. Bivariate functions which happen to be nomographic can be represented by a simple nomogram or "alignment-chart". This is done by drawing three equidistant vertical lines, and introducing a scale on each to represent $g(x)$, $\varphi(x)$, and $h(x)$. Then a straight line drawn through points labelled x and y on the first and third line will intersect the middle line at a point labelled $\varphi(g(x)+h(y))$. A special slide rule could also be designed to give the values of any specific nomographic function. By Kolmogorov's Theorem, each continuous function on the unit square is a sum of at most five nomographic functions. Thus the nomographic functions have the versatility and simplicity needed for approximation. Unfortunately, not much is known about the practical use of these functions. Most of the theory concerning them is due to R.C. Buck, who has written a series of papers on the subject. See, for example, [7] and [8].

A variant of this problem, designated (C2), arises if φ is fixed beforehand. In [22], von Golitschek has studied the discrete version of

Problem (C2) under the suposition that φ has an inverse. He established a combinatorial algorithm for finding g and h. This algorithm "converges" in a finite number of steps and has the further advantage that in each step an interval is generated which contains the number

$$\rho = \inf_{g,h} \|f - \varphi \circ (g+h)\| \quad . \tag{17}$$

Of course, when φ is fixed, the object sought is an element of a linear subspace in $C(X \times Y)$. In this sense, the problem is linear. Obviously, much remains to be done on Problem (C) and its variants.

5. Problem (D)

In Problem (D1), a continuous function f on a Cartesian product $X \times Y$ is given, and an integer n is fixed. We then seek $2n$ continuous functions $g_i \in C(X)$ and $h_i \in C(Y)$ which will minimize the norm

$$\sup_{(x,y)\in X \times Y} \left| f(x,y) - \sum_{i=1}^{n} g_i(x) h_i(y) \right| \quad . \tag{18}$$

The set of functions

$$M_n = \{ \sum_{i=1}^{n} g_i \cdot h_i : \; g_i \in C(X), \; h_i \in C(Y) \} \tag{19}$$

is a nonlinear manifold in $C(X \times Y)$, and it contains (many) linear spaces of infinite dimension.

As we shall see, this problem can be reformulated as a nasty problem in n-widths. Recall that the n-width of a set K in a Banach space E is defined by the equation

$$d_n(K) = \inf_{G} \sup_{f\in K} \operatorname{dist}(f,G) \tag{20}$$

where the infimum is taken over all n-dimensional subspaces G in E. If the infimum is attained by a specific n-dimensional subspace G, then that subspace is said to be <u>extremal</u> for K, or <u>optimal</u>.

The calculation of the distance from f to M_n (defined in Equation 19) proceeds as follows. (Here we denote by G the linear span of $g_1, \cdots, g_n$)

$$\text{(21)}\qquad \operatorname{dist}(f,M_n) = \inf_{g_i}\inf_{h_i}\Big\|f - \sum_{i=1}^{n} g_i h_i\Big\|$$

$$= \inf_{G} \operatorname{dist}[f, G \otimes C(Y)]$$

$$= \inf_{G}\sup_{y} \operatorname{dist}[f^y, G]$$

$$= d_n(\{f^y : y \in Y\}) \ .$$

In this calculation we use the tensor notation as follows:

$$\text{(22)}\qquad G \otimes C(Y) = \Big\{ \sum_{i=1}^{n} g_i \cdot h_i : h_i \in C(Y)\Big\} \ .$$

In the 3rd step of the computation the following lemma from [18] is needed.

LEMMA. Let G be a linear subspace of a Banach space E, and let X be a compact topological space. Let $C(X,E)$ be the Banach space of continuous maps $f: X \to E$ with norm $\|f\| = \sup\|f(x)\|$. Then $\operatorname{dist}[f, C(X,G)] = \sup_x \operatorname{dist}[f(x), G]$.

The determination of the n-width of $\{f^y : y \in Y\}$ as a subset of $C(X)$ is regarded as quite intractable. Also, the determination of an extremal subspace G seems to be difficult. However, the problem in Hilbert space has a classical solution, which we now outline.

In this problem, designated as (D2), we are presented with a function f in $L_2(X \times Y)$. Here X and Y are arbitrary measure spaces. For a fixed n, we seek g_i in $L_2(X)$ and h_i in $L_2(Y)$ to minimize the L_2-norm:

$$\text{(23)}\qquad \Big\{\int_X \int_Y \Big|f(x,y) - \sum_{i=1}^{n} g_i(x)h_i(y)\Big|^2\Big\}^{\frac{1}{2}} \ .$$

The solution was given by Erhard Schmidt in 1905, [53]. First define a symmetric kernel k by putting

$$\text{(24)}\qquad k(x,s) = \int_Y f(x,t)f(s,t)dt \qquad (x,s \in X) \ .$$

The eigenvalues of the kernel are the complex numbers λ for which the equation

$$\varphi(x) = \lambda \int_X k(x,s)\varphi(s)ds \tag{25}$$

has nontrivial solutions. For each eigenvalue, the solutions to Equation (25) form a finite-dimensional subspace of $L_2(X)$. The eigenvalues are positive real numbers and can be arranged as $\lambda_1 \geq \lambda_2 \geq \ldots$ with each one repeated a number of times equal to the dimension of the space of solutions of Equation (25). Next we select $g_1, g_2, \ldots, g_n$ as eigenfunctions corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. The normalization of the g's should be $\int g_i^2 = \lambda_i^{\frac{1}{2}}$. Because of the symmetry of the situation, the functions h_i can be determined in the same way, mutatis mutandis. After the g_i and h_i have been obtained, we have

$$\left\| f - \sum_{i=1}^{n} g_i h_i \right\|^2 = \|f\|^2 - \sum_{i=1}^{n} \lambda_i^{-1} . \tag{26}$$

Furthermore, the limit of each side of this equation is 0 as $n \to \infty$. Besides being accessible in Schmidt's memoir [53], these matters are discussed by M. Golomb in [27]. In Section 8 of Golomb's article, the Newton iteration is discussed as a means of determining the eigenfunctions in Schmidt's solution of the problem.

Another variant of Problem (D) arises if we simply change the norm to the L_1-norm. This we designate as Problem (D3). The paper [46] of Micchelli and Pinkus is devoted to this problem. In order to state their main theorem we need a definition: a function $f \in C([0,1]^2)$ is said to be strictly totally positive if $\det f(x_i, y_j) > 0$ whenever $0 \leq x_1 < \ldots < x_m \leq 1$, $0 \leq y_1 < \ldots < y_m \leq 1$, and $m \geq 1$.

THEOREM 10. *If f is strictly totally positive, then the minimum value of* $\int_0^1 \int_0^1 \left| f(x,y) - \sum_{i=1}^{n} u_i(x)v_i(y) \right| dxdy$ *when u_i and v_i range over* $L_1[0,1]$ *is obtained by functions of the form* $u_i(x) = f(x, \eta_i)$, $v_i(y) = \sum_{j=1}^{n} c_{ij} f(\xi_j, y)$.

In their paper, Micchelli and Pinkus specify the magical points η_i and ξ_j as well as the coefficients c_{ij}. One remarkable aspect of their work is that in this case no improvement in the quality of approximation arises from allowing u_i and v_i to range over $L_1[0,1]$; they may as well be restricted to $C[0,1]$.

The last variant of Problem (D) that we wish to discuss is designated (D4), and is obtained by fixing the functions $h_1, \ldots, h_n$ in $C(Y)$. This

restores linearity to the problem and renders it more amenable to analysis. Let H denote the subspace of $C(Y)$ generated by $h_1, \ldots, h_n$. Then our approximating subspace in $C(X \times Y)$ is

$$C(X) \otimes H = \{ \sum_{i=1}^{n} u_i \cdot h_i : u_i \in C(X) \} . \tag{27}$$

If H is not finite dimensional, then $C(X) \otimes H$ is defined as the closure in $C(X \times Y)$ of the set of all finite sums $\sum u_i v_i$ with $u_i \in C(X)$ and $v_i \in H$.

A number of general theorems about this problem are contained in [18]. Usually, H need not be finite-dimensional. As examples of the results, here are several:

THEOREM 11. *If there exists a continuous proximity map ("continuous selection for the set-valued metric projection") from $C(Y)$ onto H, then $C(X) \otimes H$ is proximinal in $C(X \times Y)$.*

THEOREM 12. *Let H be any subspace of $C(Y)$, let $f \in C(X \times Y)$, and let $g \in C(X) \otimes H$. In order that g be a best approximation to f it is necessary and sufficient that one x-section of $f - g$ satisfy $\|f-g\| = \operatorname{dist}(f_x - g_x, H)$.*

Suitable examples show that the hypothesis of Theorem 11 cannot be dropped, even if H is one-dimensional. Many of the results in [18], including stronger versions of Theorems 12 and 11, apply to the more general Banach spaces $C(X,E)$ described previously.

(12)

6. Problem (E)

Here we understand that functions h_i and g_i are fixed in $C(Y)$ and $C(X)$ respectively, and we seek continuous functions u_i and v_i to minimize

$$\sup_x \sup_y \left| f(x,y) - \sum_{i=1}^{n} u_i(x) h_i(y) - \sum_{j=1}^{m} v_i(y) g_i(x) \right| . \tag{28}$$

This is Problem (E1). If we use H and G to denote the subspaces in $C(Y)$ and $C(X)$ generated by $g_1, \ldots, g_n$ and $h_1, \ldots, h_m$, then our approximating subspace is

$$W = C(X) \otimes H + G \otimes C(Y) . \tag{29}$$

This problem is the direct descendant of the original Diliberto-Straus

problem. It retains linearity, and offers tremendous flexibility for approximation. Since the elements of W are separable functions, i.e. of the form $\sum_{i=1}^{N} a_i(x)b_i(y)$, they are ideally suited to approximating the kernels of integral transforms.

The existence question for best approximations in Problem (E1) is not completely settled. The most general result to date is the infamous "Sitting Duck Theorem" from [51]:

THEOREM 13. If G is a finite-dimensional subspace in $C(X)$ having a continuous proximity map, and if H is a finite-dimensional subspace in $C(Y)$ having a Lipschitzian proximity map, then W (as in Equation (29)) is proximinal.

Although these hypotheses may eventually turn out to be unnecessarily stringent, examples can be given to show that they cannot be dropped altogether. Thus, in [18], an example in which $H = 0$ and G is of dimension 1 is displayed, with W not proximinal. Examples of spaces having Lipschitzian proximity maps are constructed in [52].

A related result from [41] also has an asymmetrical flavor. In it, we assume that each of the compact spaces has a Borel measure defined on it such that nonvoid open sets receive nonzero, noninfinite measure. The theorem then is as follows, in a reformulation suggested by L. Sulley.

THEOREM 14. Let G and H be finite-dimensional subspaces in $C(X)$ and $C(Y)$ respectively. Assume the existence of continuous proximity maps

$$A: L_\infty(X) \to G \qquad B: C(Y) \to H .$$

Then $G \otimes C(Y) + C(X) \otimes H$ is proximinal in $C(X \times Y)$.

Let us denote by $\ell_\infty(S)$ the Banach space of all bounded real functions on a set S. There are some theorems about the existence of best approximations in $\ell_\infty(X \times Y)$ by elements of the subspace

$$Z = \ell_\infty(X) \otimes H + G \otimes \ell_\infty(Y) . \tag{30}$$

It turns out that Z is proximinal if G and H are arbitrary finite-dimensional subspaces in $\ell_\infty(X)$ and $\ell_\infty(Y)$ respectively. This is established in [26]. Now an interesting phenomenon occurs. If $f \in C(X \times Y)$, $G \subset C(X)$, and $H \subset C(Y)$, then the distance from f to Z is not less than the distance from f to W. Thus no improvement in the quality of

approximation can be achieved by allowing discontinuous coefficient functions. This is proved also in [26].

There is some reason to believe that in the situation just described some of the best approximations to f in Z will be continuous on the interior of $X \times Y$. One theorem to this effect is contained in [26], but the general case appears to be rather difficult.

The problem of characterizing best approximations in Problem (E1) has received some attention. The theorem of Havinson for Problem (B1) has been mentioned previously. It could have been given this equivalent form:

THEOREM 15. *In order that the pair* (g,h) *solve the minimization problem*

$$\min_{g \in C(X)} \min_{h \in C(Y)} \sup_x \sup_y |f(x,y)-g(x)-h(y)|$$

it is necessary and sufficient that there exist functionals φ_n *in* $C(X \times Y)^*$ *such that* $\|\varphi_n\| = 1$, *each* φ_n *has finite support, each* φ_n *annihilates functions in* $C(X)+C(Y)$, *and* $\varphi_n(f-g-h) \to \|f-g-h\|$.

The functionals φ_n referred to in this theorem correspond roughly to finite initial segments of an alternant $\{p_1, \ldots, p_n\}$ in Havinson's Theorem, the correspondence being given by the equation $\varphi_n(F) \approx (1/n) \times \sum_{i=1}^{n} (-1)^i F(p_i)$, for any $F \in C(X \times Y)$.

The generalization of Theorem 14 for the subspace W (in Equation (29)) appears in [43], and states:

THEOREM 16. *In order that an element* w *in* W *be a best approximation to a function* f *in* $C(X \times Y)$ *it is necessary and sufficient that there exist functionals* φ_n *in* $C(X \times Y)^*$ *such that* $\|\varphi_n\| = 1$, *each* φ_n *has finite support, each* φ_n *annihilates* W, *and* $\varphi_n(f-w) \to \|f-w\|$.

If we possess continuous proximity maps $P: C(X) \to G$ and $Q: C(Y) \to H$, then a generalized Diliberto-Straus algorithm can be defined. In order to do this, first extend P and Q to operate on $C(X \times Y)$ by writing

$$(\bar{P}f)(x,y) = (Pf^y)(x) \tag{31}$$

$$(\bar{Q}f)(x,y) = (Qf_x)(y) . \tag{32}$$

Thus $\bar{P}$ operates on f as a parametrized family of functions of x. Actually then $\bar{P}$ is a proximity map of $C(X \times Y)$ into $G \otimes C(Y)$. Similar

remarks apply to $\overline{Q}$. Our algorithm then is: $f_0 = f$, $f_{n+1} = f_n - A_n f_n$, where $A_n = \overline{P}_n$ for even n and $A_n = \overline{Q}_n$ for odd n.

N. Dyn was the first to give an example in which this algorithm fails [14]. Her example involves $X = Y = [0,1]$, $G = \pi_0$, and $H = \pi_1$. (We use π_n as the space of polynomials of degree at most n.) In [25] this negative result is generalized to arbitrary Haar subspaces, assuming only that $\dim G \geq 1$ and $\dim H \geq 2$. If the hypothesis of Haar subspaces is dropped, then examples can be given (with G and H having arbitrary dimensionality) in which the algorithm is effective. But these subspaces are constructed solely for this purpose and are of little practical importance [52].

Although an existence theorem of wide generality and a method of constructing best approximations are both lacking for Problem (E1), nevertheless, very practical methods exist for producing good approximations. Here we refer to the linear blending projections which W.J. Gordon developed in [28] and in many later papers. The theory is both powerful and elegant. Starting with linear projections

$$P: C(X) \twoheadrightarrow G \qquad G: C(Y) \twoheadrightarrow H \tag{33}$$

we extend these to $\overline{P}$ and $\overline{Q}$ by Equations (31) and (32).

THEOREM 17. The map $B = \overline{P} + \overline{Q} - \overline{P}\,\overline{Q}$ is a linear projection of $C(X \times Y)$ onto W (as in Equation (29)). Furthermore,

$$\|I-B\| \leq \|I-P\|\|I-Q\| . \tag{34}$$

The importance of the inequality (34) is that $\|I-B\|$ is the proper measure of how well B serves as an approximation operator. Similar remarks apply to P and Q. The essential estimate needed here is

$$\|f-Bf\| \leq \|I-B\|\operatorname{dist}(f,W) . \tag{35}$$

(It depends on B being linear and idempotent.) This inequality shows that Bf is worse than the best approximation by a factor of $\|I-B\|$, at most.

The projection B defined in Theorem 17 is the Boolean sum of $\overline{P}$ and $\overline{Q}$, often written $B = \overline{P} \oplus \overline{Q}$. It is also referred to sometimes as a blending operator. One practical choice for B is obtained by using orthogonal projections for P and Q. Thus if $\{g_1, \ldots, g_n\}$ is a basis for G chosen to be orthonormal with respect to any convenient inner product, say

(36) $$\langle u,v\rangle = \int_S u(s)v(s)\rho(s)ds$$

then $\overline{P}$ would be defined by

(37) $$(\overline{P}f)(s,t) = \sum_{i=1}^{n} \langle f^t, g_i\rangle g_i(s) \ .$$

A similar equation (or some completely different projection) could be used for $\overline{Q}$. A concise summary of blending operator theory is in [29].

Because of the flexibility and simplicity of the blending operators, one can regard the problem of approximation from the subspace W as <u>solved</u> for all practical purposes. Especially good choices for P and Q, in the case that $X = Y = [-1,1]$ and $G = \Pi_{n-1}$, $H = \Pi_{m-1}$, are, of course, interpolation operators with the nodes at the Tchebycheff abscissas (i.e. roots of T_n or extrema of T_{n-1}). It should be observed that the Jameson-Pinkus projection defined by Equation (16) is the Boolean sum of two minimal projections.

Among the many possible variants of Problem (E), we shall mention two, viz., the L_1 and L_2 versions. In Problem (E2), we fix $f \in L_2(X \times Y)$, $g_i \in L_2(X)$, and $h_i \in L_2(Y)$. Here X and Y can be arbitrary σ-finite measure spaces. We seek functions $u_i \in L_2(X)$ and $v_i \in L_2(Y)$ so as to minimize the expression

(38) $$\Big\| f - \sum_{i=1}^{n} u_i h_i - \sum_{i=1}^{m} v_i g_i \Big\|_2^2$$

$$= \int\int \Big| f(x,y) - \sum_{i=1}^{n} u_i(x)h_i(y) - \sum_{i=1}^{m} v_i(y)g_i(x) \Big|^2 dxdy \ .$$

The solution can be given at once: form the Boolean sum $B = \overline{P} \oplus \overline{Q}$, where P is the <u>orthogonal</u> projection of $L_2(X)$ onto the subspace G spanned by $g_1, \ldots, g_n$, and $\overline{P}$ is the extension defined by Equation (31). The projections Q and $\overline{Q}$ are similar. One proves readily that B is the orthogonal projection onto $G \otimes L_2(Y) + L_2(X) \otimes H$. Here we have a specific instance of the following two theorems from [51].

THEOREM 18. <u>If</u> G <u>and</u> H <u>are complemented subspaces in Banach spaces</u> E <u>and</u> F <u>respectively, and if</u> α <u>is any uniform cross-norm, then</u> $G \otimes_\alpha F + E \otimes_\alpha H$ <u>is complemented in</u> $E \otimes_\alpha F$.

THEOREM 19. <u>If</u> P <u>and</u> Q <u>are linear proximity maps defined on a Banach space</u> E, <u>then the same is true of</u> $P+Q-PQ$, <u>provided that</u> $PQP = QP$.

It should be noted that a practical method for solving the uniform approximation problem (E1) consists in replacing it by the mean-square approximation problem (E2). There is even the possibility of adjusting the measure in such a way that the L_2-solution becomes the L_∞-solution. This idea goes back to Lawson's dissertation [38].

The final variant of Problem (E) is designated (E3) and involves the L_1-metric. Thus the natural setting is $f \in L_1(X \times Y)$, $g_i \in L_1(X)$, and $h_i \in L_1(Y)$. The following theorem has been proved by Holland, Light, and Sulley [33].

THEOREM 20. *If the measure spaces* X *and* Y *have finite measure, and if* G *and* H *are finite-dimensional subspaces of* $L_1(X)$ *and* $L_1(Y)$, *respectively, then the subspace* $G \otimes L_1(Y) + L_1(X) \otimes H$ *is proximinal in* $L_1(X \times Y)$.

Prior to the work just cited, the best result along these lines required $f \in L_\infty(X \times Y)$, $G \subset L_\infty(X)$, and $H \subset L_\infty(Y)$. See [40]. Proofs of these results employ the elegant "measurable selection theorem" of Kuratowski and Ryll-Nardzewski [35]. See also [50].

References

1. Amir, D. and Z. Ziegler, Relative Chebyshev centers in normed linear spaces, J. Approximation Theory 29 (1980), 235-252.

2. Atlestam, B. and F. Sullivan, Iteration with best approximation operators, Rev. Roumaine Math. Pures Appl. 21 (1976), 125-131.

3. Aumann, G., Uber approximative Nomographie, II, Bayer. Akad. Wiss. Math.-Nat. Kl. S.-B. (1959), 103-109.

4. Bank, R. E., An automatic scaling procedure for a D'Yakanov-Gunn iteration scheme, Linear Algebra and its Applications, to appear.

5. Birkhoff, G., The algebra of multivariate interpolation. in: Constructive Approaches to Mathematical Modesl, ed. by C. V. Coffman and G. J. Fix, Academic Press (1979).

6. Brown, A. L., Finite rank approximations to integral operators which satisfy certain total positivity conditions., J. Approximation Theory 34 (1982), 42-90.

7. Buck, R. C., Approximate functional complexity, Bull. Amer. Math. Soc. 81 (1975), 1112-1114.

8. _____, Approximation theory and functional equations, I and II, J. Approximation Theory 5 (1972), 228-237; 9 (1973), 121-125.

9. Delvos, F. H., and H. Posdorf, N-th order blending, pp. 53-64 in: Constructive Theory of Functions of Several Variables, Lecture Notes in Math. 571, Springer-Verlag 1976.

10. _____, On optimal tensor product approximation, J. Approximation Theory 18 (1976), 99-107.

11. Deutsch, F., The alternating method of von Neumann, pp. 83-96 in Multivariate Approximation Theory, ed. by W. Schempp and K. Zeller, ISNM vol. 51, Birkhauser, Basel, 1979.

12. Deutsch, F., J. Mach, and K. Saatkamp, Approximation by finite rank operators, J. Approximation Theory 33 (1981), 199-213.

13. Diliberto, S. P., and E. G. Straus, On the approximation of a function of several variables by the sum of functions of fewer variables, Pacific J. Math. 1 (1951), 195-210.

14. Dyn, N., A straightforward generalization of Diliberto and Straus' algorithm does not work, J. Approximation Theory 30 (1980), 247-250.

15. _____, Perfect splines ... with applications to tensor product applications and n-widths of integral operators, Report #81-16, Tel-Aviv University, April 1981.

16. Franchetti, C., On the alternating approximation method, Bollettino Unione Math. Ital. 7 (1973), 169-175.

17. Franchetti, C., and E. W. Cheney, Simultaneous approximations and restricted Chebyshev centers in function spaces, pp. 65-88 in: Approximation Theory and its Applications, Z. Ziegler, ed. Academic Press, New York, 1981.

18. _____, Best approximation problems for multivariate functions, Bolletino Unione Math. Ital. 18-B (1981), 1003-1015.

19. Franchetti, C., and S. Holland, Two extensions of the alternating algorithm of von Neumann, Report #27, Center for Approximation Theory, Texas A&M University, January 1983. To appear, Ann. Mat. Pura Appl.

20. Franchetti, C., and W. A. Light, The von Neumann algorithm in Hilbert space, Report #28, Center for Approximation Theory, Texas A&M University, January 1983.

21. Garkavi, A. L., The best possible net and the best possible cross-section of a set in a normed space, Izv. Akad. Nauk SSSR Ser. Mat. 26 (1962), 87-106, Amer. Math. Soc. Transl. (2), 39 (1964), 111-132.

22. von Golitschek, M., Approximating bivariate functions and matrices by nomographic functions, pp. 143-151 in: Quantitative Approximation, ed. by de Vore and Scherer, Academic Press, New York 1980.

23. _____, An algorithm for scaling matrices and computing the minimum cycle mean in a digraph, Numerische Math. 35 (1980), 45-55.

24. von Golitschek, M., and E. W. Cheney, On the algorithm of Diliberto and Straus for approximating bivariate functions by univariate ones, Numer. Funct. Analysis and Optimization 1 (1979), 341-363.

25. _____, Failure of the alternating method for best approximation of multivariate functions, J. Approximation Theory, to appear.

26. _____, The best approximation of bivariate functions by separable functions, Report #179, Center for Numerical Analysis, The University of Texas at Austin.

27. Golomb, M., Approximation by functions of fewer variables", pp. 275-327 in: On Numerical Approximation, ed. by R. E. Langer, University of Wisconsin Press, Madison 1959.

28. Gordon, W. J., Distributive lattices and the approximation of multivariate functions, pp. 223-277 in: Approximations with Special Emphasis on Spline Functions, ed. by I. J. Schoenberg, Academic Press, New York 1969.

29. Gordon, W. J., and E. W. Cheney, Bivariate and multivariate interpolation with noncommutative projections, pp. 381-387 in: Linear Spaces and Approximation, ed. by P. L. Butzer and B. Sz. Nagy, Birkhauser, Basel 1978.

30. Halperin, I., The product of projection operators, Acta Sci. Math. 23 (1962), 96-99.

31. Havinson, S. Ja., A Chebyshev theorem for the approximation of a function of two variables by sums of the type $\phi(x)+\psi(y)$, Izv. Akad. Nauk SSSR 33 (1969), Math. USSR Izvestija 3 (1969), 617-632.

32. Hedberg, T., The Kolmogorov superposition theorem, pp. 267-275 in: Topics in Approximation Theory, by H. S. Shapiro, Lecture Notes in Math. vol. 187, Springer-Verlag, New York 1971.

33. Holland, S. M., W. A. Light, and L. J. Sulley, On proximinality in $L_1(T \times S)$, Proc. Amer. Math. Soc. ~~to appear~~. _____

34. Jameson, G. J. O., and A. Pinkus, Positive and minimal projections in function spaces, ~~to appear in~~ J. Approximation Theory. 37 (1983), 182-195

35. Kuratowski, K., and C. Ryll-Nardzewski, A general theorem on selectors, Bull. Acad. Polon, Sci. Ser. Sci. Math. Astronom. Phys. 13 (1965), 397-403.

36. Lambert, J. M., and P. D. Milman, Restricted Chebyshev centers of bounded subsets in an arbitrary Banach space, J. Approximation Theory 26 (1979), 71-78.

37. Laurent, P. J., and P. D. Tuan, Global approximation of a compact set by elements of a convex set in a normed space, Numerische Math. 15 (1970), 137-150.

38. Lawson, C. L., Contributions to the Theory of Linear Least Maximum Approximation, Dissertation, UCLA 1961, Reprinted by Jet Propulsion Laboratory, Pasadena, CA.

39. Light, W. A., The Diliberto-Straus algorithm in $L_1(X \times Y)$, J. Approximation Theory, to appear.

40. Light, W. A., J. H. McCabe, G. M. Phillips, and E. W. Cheney, The approximation of bivariate functions by sums of univariate ones using the L_1-metric, Proc. Edinburgh Math. Soc. 25 (1982), 173-181.

41. Light, W. A., and E. W. Cheney, Some best-approximation theorems in tensor-product spaces, Math. Proc. Cambridge Phil. Soc. 89 (1981), 385-390.

42. _____, On the approximation of a bivariate function by the sum of univariate functions, J. Approximation Theory 29 (1980), 305-322.

43. _____, The characterization of best approximations in tensor product spaces, International J. of Analysis, to appear.

44. Lorentz, G. G., Approximation of Functions, Holt, Rinehart and Winston, New York 1966.

45. Mach, J., On the existence of best simultaneous approximation, J. Approximation Theory 25 (1979), 258-265.

46. Micchelli, C. A., and A. Pinkus, Best mean approximation to a 2-dimensional kernel by tensor products, Bull. Amer. Math. Soc. 83 (1977), 400-402.

47. _____, Some problems in the approximation of functions of two variables and n-widths of integral operators, J. Approximation Theory 24 (1978), 51-77.

48. von Neumann, J., Functional Operators, vol. II, Annals of Math. Studies, #22, Princeton University Press, 1950.

49. Ofman, Ju. P., Best approximation of functions of two variables by functions of the form $\phi(x)+\psi(y)$, Izv. Akad. Nauk 25 (1961), 239-252, Amer. Math. Soc. Transl. Ser. (2) 44 (1965), 12-28.

50. Parthasarathy, T., Selection Theorems and Their Applications, Springer Lecture Notes in Mathematics vol. 263 (1972).

51. Respess, J. R., and E. W. Cheney, Best approximation problems in tensor product spaces, Pacific J. Math. 102 (1982), 437-446.

52. _____, On Lipschitzian proximity maps, pp. 73-85 in: Nonlinear Analysis and Applications, ed. by S. P. Singh and J. H. Burry, Marcel Dekker, Inc., New York 1982.

53. Schmidt, E., Zur Theorie der linearen und nichtlinearen Integralgleichungen, I., Math. Annalen 63 (1907), 433-476.

54. Smith. P. W., and J. D. Ward, Restricted centers in $C(\Omega)$, Proc. Amer. Math. Soc. 48 (1975), 165-172.

55. Wagner, D. H., Survey of measurable selection theorems, SIAM J. Control and Opt., 15 (1977), 859-903.

56. Zamjatin, V. N., Cebysev centers in the space C(S), pp. 28-35 in: First Scientific Conference of Young Scholars of the Akygei", 1971 (Russian).

Department of Mathematics, The University of Texas, Austin, TX 78712

John Dennis' paper is not available here, but will also appear elsewhere.

ITPACK ON SUPERCOMPUTERS

David R. Kincaid and Thomas C. Oppe
University of Texas at Austin

ABSTRACT

ITPACK is being adapted for efficient use on supercomputers. The results of preliminary testing of the vector version of this software package are presented. More extensive modifications are being planned and are outlined in Kincaid, Oppe, and Young [1982].

INTRODUCTION

ITPACK is a package of subroutines for solving large sparse linear systems by adaptive iterative algorithms developed at the Center for Numerical Analysis of the University of Texas at Austin. (See Kincaid, Respess, Young, and Grimes [1982].) Recently, we have been adapting ITPACK 2C for efficient use on supercomputers such as the CDC Cyber 205 and CRI Cray 1. A vector version of ITPACK has been developed. The vector version of ITPACK, like the scalar version, is written in 1966 ANSI Standard Fortran with a minimum of vector syntax so that it will run efficiently on both of these computers.

ITPACK is a package of seven iterative algorithms for solving the linear system $Ax = b$ where A is a symmetric positive definite (or mildly nonsymmetric) matrix. The basic methods are the Jacobi (J), Successive Overrelaxation (SOR), Symmetric Successive Overrelaxation (SSOR), and RS method for the reduced system. Either conjugate gradient (CG) or Chebyshev (semi-iteration, SI) acceleration is applied to each of these basic methods, except for the SOR method. The seven iterative algorithms in ITPACK correspond to the following subroutines : JCG - Jacobi Conjugate Gradient, JSI - Jacobi Semi-iteration, SOR - Successive Overrelaxation, SSORCG - Symmetric Successive Overrelaxation Conjugate Gradient, SSORSI - Symmetric Successive Overrelaxation Semi- iteration, RSCG - Reduced System Conjugate Gradient, and RSSI - Reduced System Semi-iteration. Each of these modules scales matrix A to have a unit diagonal as follows

$$(D^{-1/2} A D^{-1/2})(D^{1/2} x) = (D^{-1/2} b)$$

using the diagonal matrix $D = \mathrm{diag}(A) = (d_{ii})$ which has the same diagonal elements as A. We denote this scaled system as $Au = c$ where A

is now the scaled coefficient matrix of the linear system, $u = D^{1/2} x$, and $c = D^{-1/2} b$. Each of the basic methods is derived by choosing a splitting matrix Q^{-1} such that $Q^{-1} A$ is better conditioned than the original linear system. The resulting pre-conditioned system is written as

$$Q^{-1} Au = Q^{-1} c \quad \text{or} \quad u = (I - Q^{-1} A)u + Q^{-1} c$$

which leads to the basic iterative step

$$u^{(n+1)} = Gu^{(n)} + k \tag{1}$$

where $G = I - Q^{-1} A$ and $k = Q^{-1} c$. If scaled matrix A has "Property A", it is possible to permute its rows and columns into the following form

$$A = \begin{bmatrix} I_r & H \\ K & I_b \end{bmatrix} \tag{2}$$

where I_r and I_b are identity submatrices. Such a system is obtained when the red/black ordering is used on the mesh points of a domain for an elliptic partial differential equation discretized by the 5-point finite difference equation. (For details, see Hageman and Young [1981].)

The acceleration procedures used for the basic iterative methods are either the conjugate gradient acceleration or the Chebyshev acceleration. Both can be expressed as

$$u^{(n+1)} = p_{n+1} [q_{n+1} (Gu^{(n)} + k) + (1-q_{n+1})u^{(n)}] + (1-p_{n+1})u^{(n-1)} \tag{3}$$

where p_{n+1} and q_{n+1} are acceleration parameters which can be automatically computed in ITPACK.

We are interested in the efficient computation of (3) on vector computers. Assuming $Gu^{(n)} + k$ has been computed, $u^{(n+1)}$ is a linear combination of vectors in (1) and so its calculation is clearly vectorizable. The acceleration parameters are computed using either (i) scalar arithmetic involving maximum and minimum eigenvalues of the iteration matrix G in the case of Chebyshev acceleration or (ii) dot products of intermediate work vectors in the case of conjugate gradient acceleration. The computational work needed for (i) is trivial; that for (ii) can be reduced by using fast dot product routines. Hence, only the basic iterative step (1) remains to be vectorized.

SPARSE STORAGE SCHEMES

A critical consideration in the vectorizability of (1) is the data structure used to store the matrix A. If the data structure is such that a matrix-vector product can be vectorized, then the calculation of (1) can be made very efficient for most of the ITPACK modules. The scalar version of ITPACK stores the matrix in a row-oriented structure which makes vectorization difficult. The data structure for the vector version of ITPACK follows the ELLPACK data structure which is column-oriented and thereby allows increased vectorization of the matrix-vector product. (See Rice [1982] or Boisvert [1982].)

Consider the following matrix and how it would be stored under the storage schemes used in scalar and vector ITPACK :

$$\begin{bmatrix} 4 & -1 & -2 & 0 \\ -1 & 2 & 0 & 0 \\ -2 & 0 & 6 & -3 \\ 0 & 0 & -3 & 8 \end{bmatrix}$$

Scalar ITPACK - Symmetric Storage & Nonsymmetric Storage

A = (4,-1,-2,2,6,-3,8)
JA = (1,2,3,2,3,4,4)
IA = (1,4,5,7,8)

A = (4,-1,-2,-1,2,-2,6,-3,-3,8)
JA = (1,2,3,1,2,1,3,4,3,4)
IA = (1,4,6,9,11)

Vector ITPACK - Sparse Storage Scheme

$$\text{COEF} = \begin{bmatrix} 4 & -1 & -2 \\ 2 & -1 & 0 \\ 6 & -2 & -3 \\ 8 & -3 & 0 \end{bmatrix} \qquad \text{JCOEF} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 3 & 1 & 4 \\ 4 & 3 & 1 \end{bmatrix}$$

In scalar ITPACK, the array A contains the nonzeros of the matrix entered by rows (in symmetric storage only those on or above the main diagonal are stored), JA contains the column numbers of corresponding elements in A, and IA contains pointers into A for the first nonzero of each row. In vector ITPACK, the array COEF contains the nonzeros of the matrix with the diagonal entry in column one and with possible zero fill for short rows, JCOEF contains the column numbers for corresponding elements in COEF with possible unit fill for short rows. The elements in COEF may be automatically moved around depending on the iterative method being used in order to increase vectorization. (See method sections.)

For the COEF-JCOEF data structure, all nonzeros of the coefficient matrix are stored even if it is symmetric. This is done because vectorization is difficult if only half of the nonzeros are present. Still the user should inform ITPACK if the matrix is known to be symmetric since the iterative algorithms are slightly different for symmetric and nonsymmetric systems.

CODE SEGMENTS

Now consider possible code implementations for a matrix-vector product with these storage schemes

```
      .
      .
      .
C
C ... IA-JA-A Symmetric Storage Scheme
C
      DO 3 I = 1,N
    3 C(I) = 0.E0
      DO 10 I = 1,N
      JBGN = IA(I)
      JEND = IA(I+1) - 1
      IF (JEND .LT. JBGN) GO TO 10
      SUM = C(I)
      BI = B(I)
      DO 5 J = JBGN,JEND
         JCOL = JA(J)
         SUM = SUM + A(J)*B(JCOL)
         C(JCOL) = C(JCOL) + A(J)*BI
    5    CONTINUE
      C(I) = SUM
   10 CONTINUE
      .
      .
      .
```

```
C
C ... IA-JA-A Nonsymmetric Storage Scheme
C
      DO 20 I = 1,N
      JBGN = IA(I)
      JEND = IA(I+1) - 1
      SUM = 0.E0
      IF (JEND .LT. JBGN) GO TO 15
      DO 13 J = JBGN,JEND
         JCOL = JA(J)
         SUM = SUM + A(J)*B(JCOL)
 13      CONTINUE
 15   C(I) = SUM
 20   CONTINUE
      .
      .
      .
C
C ... COEF-JCOEF Storage Scheme
C
      DO 33 I = 1,N
      C(I) = 0.E0
 33   CONTINUE
      DO 30 J = 1,MAXNZ
         CALL VGATHR (N,B,JCOEF(1,J),WORK)
         DO 35 I = 1,N
            C(I) = C(I) + COEF(I,J)*WORK(I)
 35         CONTINUE
 30    CONTINUE
      .
      .
      .
      SUBROUTINE VGATHR (N,A,IA,B)
C
C ... VGATHR GATHERS ELEMENTS FROM ARRAY A ACCORDING TO INDEX LIST IA
C ... AND PLACES THEM INTO CONSECUTIVE LOCATIONS IN ARRAY B
C
      INTEGER IA(1)
      REAL    A(1),B(1)
      DO 10 I = 1,N
      IJ = IA(I)
      B(I) = A(IJ)
10    CONTINUE
      RETURN
      END
```

Here MAXNZ is the maximum number of nonzeros per row and WORK is a work vector of length N.

While the code in the first code segment is fine for scalar computers,

i; is entirely too complicated for vectorization on supercomputers. In the second example code segment, indirect addressing prevents DO loop 13 from vectorizing. Another drawback is the inner-product algorithm itself. It computes N dot products with short vectors (a maximum of MAXNZ in length). In general, vector computers are more efficient when working with long vectors. In the third example code segment, DO loops 33 and 35 vectorize. Also, both the Cyber 205 and Cray 1 have fast gathering instructions which can replace subroutine VGATHR. This algorithm is similar to the outer-product algorithm for a matrix-vector product. It performs MAXNZ gathering operations, MAXNZ vector multiplies, MAXNZ vector additions, and MAXNZ+1 vector assignments. Note that the price for avoiding indirect addressing is the need for additional working storage of length N. Using the COEF-JCOEF data structure, the basic iterative step was vectorized or partially vectorized for all of the ITPACK modules with a significant improvement in performance.

We now consider the several basic iterative methods and their potential for vectorization.

JACOBI METHOD

For scaled A, the splitting matrix is $Q = I$ so the basic Jacobi iterative step is

$$u^{(n+1)} = (I - Q^{-1}A)\,u^{(n)} + Q^{-1}c \quad \text{or} \quad u^{(n+1)} = (I - A)u^{(n)} + c$$

This is a simple matrix-vector product which vectorizes with the COEF-JCOEF data structure.

SOR METHOD

First, consider the Gauss-Seidel method. The scaled matrix is $A = I - L - U$ where L and U are respectively strictly lower and upper triangular matrices. Let $Q = I - L$ be the splitting matrix. This leads to the iteration step

$$u^{(n+1)} = Lu^{(n+1)} + Uu^{(n)} + c. \tag{4}$$

The basic SOR iterative step is derived from the Gauss-Seidel iteration by introducing an acceleration parameter w ($0 < w < 2$).

$$u^{(n+1)} = w(Lu^{(n+1)} + Uu^{(n)} + c) + (1-w)u^{(n)} \quad \text{or}$$

$$(I - wL)u^{(n+1)} = w(Uu^{(n)} + c) + (1-w)u^{(n)}$$

The splitting matrix is $Q = (I - wL)/w$. It is possible to vectorize the right-hand-side calculation by segregating L and U in the COEF-JCOEF data structure as follows

(5) COEF =

d_{11}	nonzeros of U	nonzeros of L
.	.	.
.	.	.
d_{nn}	.	.

Thus, a given column of COEF contains only diagonal entries, only U nonzero entries, or only L nonzero entries. Once the right-hand-side is calculated, $u^{(n+1)}$ is found by a forward elimination. This calculation is recursive in nature and is not easily vectorizable. (See, for example, Kincaid, Oppe, and Young [1982].)

When the red/black ordering is used, the triangular matrices L and U from (2) are of the form

(6) $$L = \begin{bmatrix} 0 & 0 \\ -K & 0 \end{bmatrix} \qquad U = \begin{bmatrix} 0 & -H \\ 0 & 0 \end{bmatrix}$$

The SOR method can be written as

$$u_r^{(n+1)} = w(-Hu_b^{(n)} + c_r) + (1-w)u_r^{(n)}$$

$$u_b^{(n+1)} = w(-Ku_r^{(n+1)} + c_b) + (1-w)u_b^{(n)}$$

Here the unknowns are partitioned into those corresponding to red and black mesh points denoted u_r and u_b, respectively. The COEF-JCOEF data structure is

(7) COEF =

$$\left[\begin{array}{c|c} d_{11} & \text{nonzeros} \\ \cdot & \text{of} \\ \cdot & H \\ \cdot & \cdot \\ \hline \cdot & \text{nonzeros} \\ \cdot & \text{of} \\ \cdot & K \\ d_{nn} & \cdot \end{array}\right]$$

Clearly, the SOR method vectorizes with the red/black ordering.

SSOR METHOD

Similar remarks to those for the SOR method apply to the SSOR method with splitting matrix $Q = (I-wL)(I-wU)/(w(2-w))$. The SSOR basic iterative step is

$$u^{(n+1/2)} = w(Lu^{(n+1/2)} + Uu^{(n)} + c) + (1-w)u^{(n)}$$

$$u^{(n+1)} = w(Uu^{(n+1)} + Lu^{(n+1/2)} + c) + (1-w)u^{(n+1/2)}$$

The basic iterative step is partially vectorizable using the segregated data structure (5) and is fully vectorizable under red/black ordering. With red/black ordering, the optimum relaxation factor for the SSOR method is unity so that it reduces to the Symmetric Gauss Seidel (SGS) method. The SGS method requires more iterations for convergence than the SSOR method with a good relaxation factor. For scalar computers, the SSOR method loses its effectiveness with red/black ordering; however, this may not be true for vector computers. (See section on numerical results.)

RS METHOD

The RS method for the reduced system requires that the scaled matrix A be of the form (2). Using (4) and (6), the reduced system is $(I - KH)u_b = -Kc_r + c_b$. The splitting matrix is $Q = I_b$ and the iterative statement can be written as

$$u_r^{(n+1)} = -Hu_b^{(n)} + c_r$$

$$u_b^{(n+1)} = -Ku_r^{(n+1)} + c_b$$

This is the basic iterative step for the RS method. The COEF-JCOEF data structure used is given by (7). Clearly, it also is vectorizable since only matrix-vector products are computed.

NUMERICAL RESULTS

For purposes of comparing the scalar and vector versions of ITPACK, the following sample problem was considered: Solve the elliptic partial differential equation $u_{xx} + 2u_{yy} = 0$ inside the unit square with $u = 1 + xy$ on the boundary. Using the 5-point finite difference approximation and a mesh size of $h = 1/20$, the resulting linear system has 361 unknowns and a maximum of five nonzeros per row. This problem was solved using both the scalar and vector version of ITPACK with two ordering and two storage schemes. The timing routine recorded the iteration time and the total time which included the iteration time as well as time spent scaling the matrix and permuting it if red/black ordering was used.

For the scalar version of ITPACK, the natural and red/black orderings were used with the symmetric and nonsymmetric IA-JA-A storage scheme. This was done also for the vector version of ITPACK but with the COEF-JCOEF storage scheme. Since both versions were to be run on the Cyber 205 and Cray 1 computers, DO loops which had been unrolled in the original ITPACK for scalar optimization were re-rolled to increase vectorization. Table 1 contains the vector ITPACK results and the scalar ITPACK results in parentheses using the Cyber 205. In this computer, "large pages" were used and "unsafe vectorization", i.e., compiler options BOUV. (For additional details, see Mossberg [1981] or Kincaid, Oppe, and Young [1982].) On the Cyber 205, the gather and scatter routines Q8VGATHR and Q8VSCATR were used in the vector version. Moreover, the dot product routine Q8SDOT was used in both versions. Table 2 contains the vector ITPACK results and the scalar ITPACK results in parentheses using the Cray 1. The Cray gather/scatter calls were used in the vector version. Moreover, the Cray BLAS routines were used in both versions.

A comparison of the numerical results shows a considerable improvement in performance of the vector version of ITPACK from the scalar version on both supercomputers. For example in the vector version, the total time spent in each method is not significantly greater than the iteration time while in the scalar version this was not the case. This

results from the fact that scaling and permuting the system is easily vectorizable with the COEF-JCOEF data structure but not with the A-IA-JA data structure. Note also that the vector version has slightly greater workspace requirements, namely, an additional N for all methods except for JSI and SOR which require 2N more. An interesting observation is the improved performance of the SSOR methods with the red/black ordering in the vector version of this package despite the greater number of iterations needed.

Table 1 - Cyber 205 timing results in seconds for vector ITPACK (scalar ITPACK results in parentheses)

METHOD	ITERATION TIME		TOTAL TIME		WORKSPACE USED		NO. OF ITERN.
	Natural Ordering with Symmetric Storage						
JCG	.019	(.076)	.023	(.083)	1927	(1566)	61
JSI	.030	(.138)	.034	(.145)	1444	(722)	108
SOR	.073	(.142)	.083	(.150)	1083	(361)	72
SSORCG	.043	(.088)	.053	(.096)	2561	(2200)	17
SSORSI	.052	(.097)	.062	(.105)	2166	(1805)	23
RSCG	.010	(.047)	.018	(.103)	1324	(963)	31
RSSI	.018	(.092)	.027	(.148)	902	(541)	60
	Natural Ordering with Nonsymmetric Storage						
JCG	.050	(.137)	.054	(.149)	2053	(1692)	62
JSI	.030	(.194)	.034	(.205)	1444	(722)	108
SOR	.073	(.133)	.083	(.144)	1083	(361)	72
SSORCG	.045	(.085)	.055	(.097)	2595	(2234)	17
SSORSI	.052	(.113)	.062	(.124)	2166	(1805)	23
RSCG	.020	(.061)	.028	(.150)	1386	(1025)	31
RSSI	.018	(.100)	.027	(.189)	902	(541)	60
	Red/Black Ordering with Symmetric Storage						
JCG	.018	(.071)	.027	(.127)	1927	(1566)	61
JSI	.030	(.130)	.038	(.186)	1444	(722)	108
SOR	.018	(.124)	.025	(.180)	1083	(361)	65
SSORCG	.030	(.195)	.039	(.251)	2611	(2250)	42
SSORSI	.040	(.205)	.048	(.261)	2166	(1805)	69
RSCG	.011	(.047)	.018	(.104)	1324	(963)	31
RSSI	.018	(.091)	.027	(.148)	902	(541)	60
	Red/Black Ordering with Nonsymmetric Storage						
JCG	.049	(.137)	.058	(.225)	2053	(1692)	62
JSI	.030	(.194)	.038	(.282)	1444	(722)	108
SOR	.018	(.119)	.025	(.207)	1083	(361)	65
SSORCG	.045	(.229)	.053	(.317)	2715	(2254)	47
SSORSI	.041	(.295)	.049	(.383)	2166	(1805)	69
RSCG	.021	(.061)	.028	(.149)	1386	(1025)	31
RSSI	.018	(.100)	.027	(.189)	902	(541)	60

Table 2 - Cray 1 timing results in seconds for vector ITPACK (scalar ITPACK results in parentheses)

METHOD	ITERATION TIME		TOTAL TIME		WORKSPACE USED		NO. OF ITERN.	
	Natural Ordering with Symmetric Storage							
JCG	.025	(.089)	.029	(.097)	1927	(1568)	61	(62)
JSI	.044	(.157)	.048	(.164)	1444	(722)	108	
SOR	.104	(.139)	.114	(.148)	1083	(361)	72	
SSORCG	.059	(.086)	.070	(.094)	2561	(2200)	17	
SSORSI	.073	(.087)	.083	(.096)	2166	(1805)	23	
RSCG	.013	(.053)	.022	(.109)	1324	(963)	31	
RSSI	.024	(.103)	.033	(.160)	902	(541)	60	
	Natural Ordering with Nonsymmetric Storage							
JCG	.046	(.132)	.051	(.143)	2053	(1692)	62	
JSI	.043	(.201)	.047	(.212)	1444	(722)	108	
SOR	.104	(.129)	.115	(.140)	1083	(361)	72	
SSORCG	.061	(.082)	.071	(.093)	2595	(2234)	17	
SSORSI	.072	(.099)	.083	(.111)	2166	(1805)	23	
RSCG	.020	(.063)	.029	(.150)	1386	(1025)	31	
RSSI	.024	(.109)	.033	(.197)	902	(541)	60	
	Red/Black Ordering with Symmetric Storage							
JCG	.025	(.077)	.034	(.131)	1927	(1568)	61	(62)
JSI	.044	(.133)	.053	(.188)	1444	(722)	108	
SOR	.025	(.116)	.034	(.171)	1083	(361)	65	
SSORCG	.041	(.210)	.050	(.266)	2611	(2260)	42	(47)
SSORSI	.059	(.190)	.068	(.245)	2166	(1805)	69	(57)
RSCG	.013	(.053)	.022	(.110)	1324	(963)	31	
RSSI	.024	(.104)	.033	(.159)	902	(541)	60	
	Red/Black Ordering with Nonsymmetric Storage							
JCG	.046	(.132)	.055	(.218)	2053	(1692)	62	
JSI	.043	(.202)	.052	(.288)	1444	(722)	108	
SOR	.025	(.117)	.034	(.203)	1083	(361)	65	
SSORCG	.056	(.218)	.065	(.305)	2715	(2354)	47	
SSORSI	.060	(.246)	.069	(.332)	2166	(1805)	69	(59)
RSCG	.019	(.062)	.028	(.150)	1386	(1025)	31	
RSSI	.024	(.110)	.033	(.197)	902	(541)	60	

ACKNOWLEDGEMENTS

The authors wish to thank William M. Coughran and Jessica Hodgins of Bell Laboratories for obtaining the numerical results for the Cray 1 computer. The Cyber 205 numerical results were obtained using CyberNet with assistance from Bjorn Mossberg of Control Data Corporation. This

work was supported in part by the Control Data Corporation grant 81T01, by the National Science Foundation grant MCS-79-19829, and by the Department of Energy grant DE-AS05-81ER10954.

REFERENCES

Boisvert, R. F. [1982]. ELLPACK contributor's guide, National Bureau of Standards, Scientific Computing Division, Washington, D. C. 20234.

CDC Reference Manual [1979]. "CDC CYBER 200 Fortran Language 1.4" Publ. 60457040, Control Data Corp., 215 Moffett Park Dr., Sunnyvale, Calif. 94086.

Grimes, R. G., Kincaid, D. R., Young, D. M. [1979]. ITPACK 2.0 user's guide. CNA-150, Center for Numerical Analysis, Univ. of Texas, Austin, Tex., 78712.

Hageman, L. A., and Young, D. M. [1981]. "Applied Iterative Methods." Academic Press, New York.

Kincaid, D. R., Oppe, T., Young, D. M. [1982]. Adapting ITPACK routines for use on a vector computer. CNA-177, Center for Numerical Analysis, Univ. of Texas, Austin, Tex., 78712.

Kincaid, D. R., Respess, J. R., Young, D. M., and Grimes, R. G. [1982]. ITPACK 2C: A Fortran package for solving large sparse linear systems by adaptive accelerated iterative methods. ACM Trans. Math. Softw. 8, 3.

Kincaid, D. R., and Young, D. M. [1979]. Survey of iterative methods. In Encyclopedia of Computer Sciences and Technology 13, J. Belzer, A. Holzman, and A. Kent (Eds.), Marcel Dekker, New York, 354-391.

Lawson, C. L., Hanson, R. J., Kincaid, D. R., and Krogh, F. T. [1979]. Basic linear algebra subprograms for Fortran usage. ACM Trans. Math. Softw. 5, 3, 308-323.

Mossberg, B. [1981]. An informal approach to number crunching on the Cyber 203/205. Control Data Corp., 215 Moffett Park Dr., Sunnyvale, Calif. 94086.

Rice, J. R. [1982]. ELLPACK user's guide. CSD-TR-372, Computer Science Dept., Purdue University, West Lafayette, Ind. 47907.

Young, D. M. [1971]. "Iterative Solution of Large Linear Systems." Academic Press, New York.

Higher Order Methods for Solving Algebraic Equations

Beny Neta
Department of Mathematics
Texas Tech University
Lubbock, Tx 79409

Abstract

The first part of this talk will review higher order methods for obtaining a simple zero of a nonlinear function. The order and relative efficiency of the methods are discussed. In case of multiple zeros a new method of order three is developed. The method requires 2 evaluations of the function and one evaluation of the derivative.

The second part is devoted to higher order methods for the solution of nonlinear systems of algebraic equations. Methods of order $1+\sqrt{2}$ and 4 for regular systems will be described. Higher order methods for systems having singular Jacobian will be reviewed (results due to Rall, Reddien, Decker & Kelley, Griewank, Weber & Werner and Neta & Victory).

§1. Introduction

In this talk we review higher order methods for obtaining a simple zero of a nonlinear function. The order and efficiency of these methods will be discussed. We also give a table comparing the efficiency of known methods. An extensive bibliography for the numerical solution of a nonlinear algebraic equation (having simple or multiple roots) will be given elsewhere.

We also review higher order methods for the solution of nonlinear algebraic systems.

In the next section we give some definition. The third section will give results concerning the numerical solution of non-linear equation. The last section will be devoted to systems of

equations.

§2. Preliminaries

The nonlinear equation cosidered is

$$f(x) = 0, \tag{1}$$

and the system of equations

$$\underline{F}(\underline{x}) = 0. \tag{2}$$

We denote a solution of (1) by ξ .

Definition Let $\{x_i\}$ be a sequence converging to ξ. Let $\varepsilon_i = x_i - \xi$. If there exists a real number p and a nonzero constant C such that

$$\frac{|\varepsilon_{i+1}|}{|\varepsilon_i|^p} \to C, \tag{3}$$

then p is called the order of the sequence.

Definition. The information usage d of a scheme is defined as the number of new pieces of information required per step.

Definition. The informational efficiency EFF is defined by

$$EFF = p/d. \tag{4}$$

Definition. The efficiency index EFF* is given by

$$EFF^* = p^{1/d}. \tag{5}$$

§3. Solution of a Nonlinear Algebraic Equation

There are two classes of methods for numerically solving a nonlinear equation. The first class is called bracketing, which includes the bisection, Regula Falsi and others. The idea there is to start with an interval containing the zero and refine it. The

process gives a sequence of subintervals whose length tends to zero and all containing the zero. The other class is called fixed point and include Newton's method. Here one starts with a point x_o close to the zero ξ and contruct a sequence $\{x_i\}$ converging to ξ. For example, Newton's method

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} . \tag{6}$$

It is known that if ξ is a simple zero then the order is 2 and if ξ is a multiple zero the order of the method is 1. If one knows the multiplicity m of the root ξ then the following modification due to Schröder [39] yields a second order method

$$x_{i+1} = x_i - m \frac{f(x_i)}{f'(x_i)} . \tag{7}$$

If m is not known, see Traub [41 ,pp. 129-130] for a way to approximate it.

In this talk we consider only methods from the second class. Newton's method for computing a simple zero ξ of a nonlinear equation has been modified in a number of ways. For example, Ostrowski [29] discusses a third order method that evaluates the function f at every substep but only requires the derivative f' at every other substep

$$\begin{aligned} \omega_i &= x_i - \frac{f(x_i)}{f'(x_i)} , \\ x_{i+1} &= \omega_i - \frac{f(\omega_i)}{f'(x_i)} . \end{aligned} \tag{8}$$

He also discusses a fourth order scheme that uses the same information. King [20] has shown that there is a family of such fourth order methods,

$$\omega_i = x_i - \frac{f(x_i)}{f'(x_i)},$$

$$x_{i+1} = \omega_i - \frac{f(\omega_i)}{f'(x_i)} \frac{f(x_i)+\beta f(\omega_i)}{f(x_i)+(\beta-2) f(\omega_i)} . \qquad (9)$$

Traul [41] introduced a third order method which requires one function evaluation and two evaluations of f' per step,

$$x_{i+1} = x_i - a_1 \frac{f(x_i)}{f'(x_i)} - a_2 \frac{f(x_i)}{f'(x_i + \beta f(x_i)/f'(x_i))}. \qquad (10)$$

Popovski [30] introduced a third order one parameter family of methods requiring one evaluation of each f,f' and f". (Halley's [16] method and Cauchy's [8] method are special cases). Jarratt [17] developed a fourth order method which uses one function evaluation and two evaluation of f'. King [19] introduced a fifth order scheme requiring two evaluations of each f and f',

$$\omega_i = x_i - \frac{f(x_i)}{f'(x_i)},$$

$$x_{i+1} = \omega_i - \frac{f(\omega_i)}{f'(\omega_i)} - \frac{f^3(\omega_i)}{f'(x_i)} \frac{1}{(f(x_i)+(\beta_1-2) f(\omega_i))(f(x_i)+(\beta_2-2) f(\omega_i))}. \qquad (11)$$

Werner [46] developed a scheme of order $1+\sqrt{2}$ that requires one evaluation of each f and f'. If the root is multiple the method is of order 2. The order cannot be improved upon by using the multiplicity.

$$x_{i+1} = x_i - \frac{f(x_i)}{f'\left(\frac{x_i+y_i}{2}\right)},$$

$$y_{i+1} = x_{i+1} - \frac{f(x_{i+1})}{f'\left(\frac{x_i+y_i}{2}\right)}. \qquad (12)$$

Popovski [34] developed sixth order methods requiring two function evaluations and one evaluation of f' and f".

$$z_i = \phi(x_i),$$
$$x_{i+1} = z_i - \frac{f(z_i)}{3\frac{f(z_i)-(fx_i)}{z_i-x_i} - 2\,f'(x_i) - \frac{1}{2}(z_i-x_i)f''(x_i)}, \tag{13}$$

where the first substep is any third order one point method like Halley, Cauchy, Chebyshev [31].

Neta [23] has constructed a sixth order method that requires three function evaluations and one of f',

$$\omega_i = x_i - \frac{f(x_i)}{f'(x_i)},$$
$$z_i = \omega_i - \frac{f(\omega_i)}{f'(x_i)}\,\frac{f(x_i)+Af(\omega_i)}{f(x_i)+(A-2)f(\omega_i)}, \tag{14}$$
$$x_{i+1} = z_i - \frac{f(z_i)}{f'(x_i)}\,\frac{f(x_i)-f(\omega_i)}{f(x_i)-3f(\omega_i)}.$$

If A=-1, then the correcting factor in the last two substeps is the same. Popovski [33] has shown that this method is of order 7. Neta [26] has improved this result and constructed a method of order 10.815 using the same information. Neta [24] also developed methods of order 14 and 16 using four evaluations of f and one of f'. Let us now compare the efficiency of these methods and others.

Method	Order	Information	EFF
Neta	16	5	3.2
Neta	14	5	2.8
Neta	10.815	4	2.704
Popovski [32]	7.464	4	1.866
Muller [21]	1.839	1	1.839
Popovski	7	4	1.75
Secant [19]	1.618	1	1.618
Neta	6	4	1.5
Popovski	6	4	1.5
Ostrowski	4	3	1.333
King	4	3	1.333
Jarratt	4	3	1.333
King	5	4	1.25
Murakami [22]	5	4	1.25
Werner	$1+\sqrt{2}$	2	1.207
Ostrowski	3	3	1
Traub	3	3	1
Popovski	3	3	1
Newton	2	2	1
Steffensen [40]	2	2	1

Table 1

Method	Order	Information	EFF*
Muller	1.839	1	1.839
Neta	10.815	4	1.813
Neta	16	5	1.741
Neta	14	5	1.695
Popovski	7.464	4	1.653
Popovski	7	4	1.626
Secout	1.618	1	1.618
Ostrowski	4	3	1.587
King	4	3	1.587
Jarratt	4	3	1.587
Neta	6	4	1.565
Popovski	6	4	1.565
Werner	$1+\sqrt{2}$	2	1.554
King	5	4	1.495
Murakami	5	4	1.495
Ostrowski	3	3	1.442
Traub	3	3	1.442
Popovski	3	3	1.442
Newton	2	2	1.414
Steffeusen	2	2	1.414

Table 2

Note that this comparison is done assuming the difficulty of evaluating the function and its derivatives is the same.

What happens if the zero ξ is of multiplicity m>1. One can use the modified Newton's method suggested by Schröder [39]. One can iterate using $\frac{f}{f'}$ or $f^{(m-1)}$ or $f^{1/m}$. All these functions have a

simple zero ξ. Clearly these introduce higher order derivatives. Victory and Neta [42] have constructed a two step method of order three for approximating multiple zeros of f(x),

$$\omega_i = x_i - \frac{f(x_i)}{f'(x_i)} \quad ,$$

$$x_{i+1} = \omega_i - \frac{f(\omega_i)}{f'(x_i)} \; \frac{f(x_i) + Af(\omega_i)}{f(x_i) - Bf(\omega_i)} \quad , \tag{15}$$

where

$$\mu = \frac{m}{m-1} \quad , \tag{16}$$

$$A = \mu^{2m} - \mu^{m+1} \quad , \tag{17}$$

$$B = \frac{m-2}{m-1} \mu^m + \frac{1}{(m-1)^2} \quad . \tag{18}$$

The following two examples compare this method with Werner's (12) and modified Newton's (7).

Example 1 $f(x) = x^2 e^x$. (19)

This function has a double zero at zero.

Iteration	x		
Number	(12)	(15)	(15)
0	.2	.2	1.
1	.2160-1	.1487-2	.1141
2	.2379-3	.6844-9	.2904-3
3	.2830-7	-	.5103-11

Note that method (15) requires only 2 iterations and obtains 2 more digits of accuracy. Starting with x_0=1 instead of x_0=.2, method (15) needs 3 iterations but gives 4 more digits of accuarcy.

Example 2 $f(x) = 3x^4 + 8x^3 - 6x^2 - 24x + 19.$ (20)

This function has a double root at x=1.

Iteration	x		order	
Number	(7)	(15)	(7)	(15)
0	0.	0.	-	-
1	1.583333	.8904491	-	-
2	1.071987	.9998828	1.75	3.26
3	1.001386	1.0000000	1.985	-
4	1.0000001	1.0000000	2.0002	
5	1.0000000	-		
6	1.0000000	-		

§4. Solution of Nonlinear Algebraic Systems

In recent years there has been much interest in n-dimensional variations of Newton's method, secant method and other classical one dimensional iterative methods.

The algorithm in most common usage for solving the system

$$\underline{F}(\underline{x}) = 0 \tag{21}$$

is the following

i. Given $\underline{x}_i$, $\underline{F}(\underline{x}_i)$, and the Jacobian $J(\underline{x}_i)$

ii. Solve the nxn linear system

$$J(\underline{x}_i)\ \underline{\Delta}_i = -\ \underline{F}(\underline{x}_i) \text{ for } \underline{\Delta}_i \tag{22}$$

iii. Use $\underline{\Delta}_i$ and perhaps some other values of $\underline{F}$ to choose $\underline{x}_{i+1}$

iv. Evaluate $\underline{F}(\underline{x}_{i+1})$ and test for convergence. Either terminate the computation or go to next step.

v. Evaluate or approximate $J(\underline{x}_{i+1})$, set the counter to i+1 and return to step ii.

The traditional area of research on quasi-Newton methods are steps iii and v. The reason is that for real problems, evaluations of $\underline{F}$ and J dominate the cost of solution and so it is in these steps that the potential saving in greater.

Werner's [46] method (12) of order $1+\sqrt{2}$ is applicable for systems when written as

$$\begin{aligned} J\left(\frac{\underline{x}_i + \underline{y}_i}{2}\right)(\underline{x}_{i+1} - \underline{x}_i) &= -\underline{F}(\underline{x}_i), \\ J\left(\frac{\underline{x}_i + \underline{y}_i}{2}\right)(\underline{y}_{i+1} - \underline{x}_{i+1}) &= -\underline{F}(\underline{x}_{i+1}) \end{aligned} \tag{23}$$

Neta [25] has developed a fourth order method

$$\begin{aligned} J(\underline{x}_i)(\underline{\omega}_i - \underline{x}_i) &= -\underline{F}(\underline{x}_i), \\ J(\underline{x}_i)(\underline{z}_i - \underline{\omega}_i) &= -D\underline{F}(\underline{\omega}_i), \\ J(x_i)(x_{i+1} - z_i) &= -D\underline{F}(\underline{z}_i), \end{aligned} \tag{24}$$

where the diagonal matrix D has elements

$$D_{jj} = \frac{F_j(\underline{x}_i) - F_j(\underline{\omega}_i)}{F_j(\underline{x}_i) - 3F_j(\underline{\omega}_i)} \tag{25}$$

Note that J is computed and factored only once. Extensive numerical experiments with this method show that one can save at least 20% of the time to solve a system using 2 Newton substeps (fourth order). The saving is greater if the number of iterations or the dimension of the system is larger. For example, for a system of 169x169 the saving is 38%.

For systems whose Jacobian is singular there were linear and

superlinear methods developed by Griewank [14-15], Decker and Kelley [10-12] and others.

Let F be a C^3 mapping from a Banach space E into E. Let the Frechet derivative F' be singular at the solution ξ and has finite dimensional null space N and closed range X so that $E = N \oplus X$. Let P_N be a projection onto N parallel to X and let $P_x = I - P_N$. The singular set of F' near ξ may then range from a single point to a codimension one smooth manifold through ξ. The assumptions described later force the second situation to occur. Hence one can expect nonsingularity of F' only in carefully selected regions about ξ. The Newton iterates must remain in the chosen region of invertibility of F'. A set which will satisfy these requirements can be defined as follows

$$W_{\rho,\theta} = \{x \in E \mid 0 < \| x-\xi \| < \rho ,$$

$$\| P_x(x-\xi) \| \leq \theta \| P_N(x-\xi) \| \}, \ \rho > 0, \theta > 0. \tag{26}$$

Theorem (Reddien [36])

Let 1. dim N=1

2. $F''(\xi)(N,N) \cap X = \{0\}$

3. There is $c > 0$ such that $\forall \phi \in N$, $x \in N$

$$\| F''(\xi)(\phi,x) \| \geq c \|\phi\| \ \|x\|. \tag{27}$$

Then, for ρ,θ sufficiently small $[F'(x)]^{-1}$ exists for $x \in W_{\rho,\theta}$, the map

$$Gx = x-[F'(x)]^{-1}F(x) \tag{28}$$

takes $W_{\rho,\theta}$ into itself, and there is $c_1 > 0$ so that

$$\| [F'(x)]^{-1} \| \leq c_1 \| x-\xi \|^{-1}, \ \forall x \varepsilon W_{\rho,\theta}. \tag{29}$$

Moreover, if $x_o \varepsilon W_{\rho,\theta}$ and

$$x_i = Gx_{i-1}, i = 1,2,... \tag{30}$$

the sequence $\{x_i\}$ converges to ξ and

$$\| P_x(x_i-\xi) \| \leq c_2 \| x_{i-1} - \xi \|^2, \ c_2>0 \tag{31}$$

$$\lim_{i\to\infty} \frac{\| P_N(x_i - \xi) \|}{\| P_N(x_{i-1} - \xi) \|} = \frac{1}{2} \tag{32}$$

Decker and Kelley [11] generalized this result

Theorem [11]

The above conclusions hold under the following assumptions

1. dim N < ∞
2. For all z ε N,

$$B(z) = - P_N F''(\xi)(z, P_N .) \tag{33}$$

is a nonsingular map on N.

Note that if dim N=1 then the last 2 assumptions in Reddien's theorem imply the second assumption above.

Another result by Decker and Kelley [12] shows that one can obtain a quadratically convergent method but the function F and its Frechet derivative should be evaluated twice each step.

Theorem [12]

Let 1. dim N=1 or 2

2. B(z) a nonsingular map on N for all nonzero zεN.
3. There is c>0 so that for all $\phi \varepsilon N$

$$\| F'''(\xi)(\phi,\phi,\phi) \| > c \| \phi \|^3. \tag{34}$$

Then, for ρ,θ sufficiently small and $x_0 \varepsilon W_{\rho,\theta}$, the sequence $\{x_i\}$ given by

$$x_{2i+1} = x_{2i} - [F'(x_{2i})]^{-1}F(x_{2i}),$$
$$x_{2i+2} = x_{2i+1} - (I+P_N)[F'(x_{2i+1})]^{-1}F(x_{2i+1}), \tag{35}$$

is contained in $W_{\rho,\theta}$ and converges to ξ. Moreover, there is $M>0$ so that

$$\| x_{2i+2} - \xi\| \leq M \|x_{2i} - \xi\|^2. \tag{36}$$

Griewank [15] has extended these results to prove convergence of the iterates even if x_0 is in a starlike domain (i.e. F' is nonsingular in part of $W_{\rho,\theta}$).

Weber and Werner [45] suggested to solve an auxiliary problem possessing an isolated solution. Let $\hat{E} = ExE\,x\mathbb{R}$ be equipped with the norm

$$\| W\| = \| W_1\|_E + \| W_2\|_E + |W_3|. \tag{37}$$

where $W = (W_1,W_2,W_3) \varepsilon \hat{E}$ and $W_1,W_2 \varepsilon E$, $W_3 \varepsilon \mathbb{R}$. The auxiliary problem is

$$G(W) = \begin{bmatrix} F(W_1) + W_3W_2 \\ F'(W_1)W_2 \\ a(W_2,W_2) - 1 \end{bmatrix} \tag{38}$$

where $a(W_2,W_2)$ is a continuous symmetric positive definite bilinear form on ExE. The solution of $G(W) = 0$ is $\zeta = (\xi,v,0)$, since we may assume without loss of generality that $a(v,v) = 1$.

Theorem [45]

Under the same hypotheses as Reddien's, then $G'(W)$ is a linear homeomorphism on $\hat{E}$ and ζ is an isolated solution of $G(W) = 0$.

Remark. If 0 is an eigenvalue of geometric multiplicity 1 but algebraic multiplicity greater than 1 one has to consider the following problem

$$H(W)=\begin{bmatrix} F'(W_1)^T F(W_1)+ W_3 W_2 \\ F'(W_1)W_2 \\ a(W_2,W_2) - 1 \end{bmatrix} . \tag{39}$$

Weber and Werner have suggested to solve the auxiliary problem using Werner's algorithm (23). Neta and Victory [27] have suggested to use Neta's fourth order algorithm (24). If one defines the efficiency of a method as the order per total number of multiplications one can show that method (24) is more efficient than (23) for all $n\geq 2$, where n is the dimension of the auxiliary problem. The efficiency of algorithm (24) is

$$e_N = \frac{4}{n(n^2+8n+n)} , \tag{40}$$

and that of (23) is

$$e_W = \frac{1+\sqrt{2}}{n^2(n+6)} . \tag{41}$$

Remark. It is not advisable to turn too early to the auxiliary problem, as was demonstrated in [45,27]. One also should check that the solution the sequence converged to is not a suprious solution, see example in [27].

REFERENCES

1. Aldrich, L.E., Solution of algebraic equations,J. Franklin Inst., 256 (1953), 59-69.

2. Altman, M., Iterative methods of higher order Bull, Acad. Sci. Polon. Sci. Ser. Math. Astr. Phys., 9 (1961), 63-68.

3. Altman, M., A generalization of Newton's method,Bull. Acad. Polon. Sci. Cl III, 3 (1955), 189-193.

4. Bateman, E.H., Halley's methods of solving equations, Amer. Math. Monthly, 45 (1938), 11-17.

5. Barzilai, J., Quasi-Newton methods converge at the golden section rate, Research Report #403, Center for Cybernetic Studies, Univ. of Texas, Austin, 1981.

6. Blackburn, J.A., Beaudoin, Y., A note on Chambers' methods, Math. Comp., 28 (1974), 573-574.

7. Brent, R.P., Winograd, S., Wolfe, P., Optimal iterative processes for root-finding, Numer. Math., 20 (1973), 327-341.

8. Cauchy, A. Sur la resolution numerique des equations alsebriques et transcendantes, C. R. Acad. Sci. Paris, 11 (1840), 829-847.

9. Cavanagh, R.C., Difference equations and iterative processes, Ph.D. Thesis, Computer Science Dept., Univ. of Maryland, College Park, 1970.

10. Decker, D.W., Kelley, C.T., Newton's method at singular points I, SIAM J. Numer. Anal. 17 (1980), 66-70.

11. Decker, D.W., Kelley, C.T., Newton's method at singular points II, SIAM J. Numer. Anal. 17 (1980), 465-471.

12. Decker, D.W., Kelley, C.T., Convergence acceleration for Newton's method at singular points, SIAM J. Numer. Anal. 19 (1982), 219-229.

13. Dennis, J.E., More, J.J., Quasi-Newton methods, motivation and theory, SIAM Rev. 19 (1977), 46-89.

14. Griewank, A.O., Starlike domains of convergence for Newton's method at singularities, Numer. Math. 35 (1980), 95-111.

15. Griewank, A.O., Osborne, M.R., Analysis of Newton's method at irregular singularities (1982) to appear.

16. Halley, E., A new, exact and easy method of finding the roots of equations generally and that without any previous reduction, Phil. Trans. Roy. Soc. London, 18 (1694), 136-148.

17. Jarratt, P., Some fourth order multipoint methods for solving equations, Math. Comp., 20 (1966), 434-437.

18. Johnson, L.W., Riess, R.D., Numerical Analysis. Addison-Wesley Pub. Co., Mass., 1977.

19. King, R.F., A fifth-order family of modified Newton methods, BIT, 11 (1971), 409-412.

20. King, R.F., A family of fourth-order methods for nonlinear equations, SIAM J. Numer. Anal., 10 (1973), 876-879.

21. Muller, D.E., A method for solving algebraic equations using an automatic computer, Math. Comp., 10 (1956), 208-215.

22. Murakami, T., Some fifth-order multipoint iterative formulae for solving equations, J. of Information Processing, 1 (1978), 138-139.

23. Neta, B., A sixth-order family of methods for nonlinear equations, Intern. J. Computer Math., 7 (1979), 157-161.

24. Neta, B., On a family of multipoint methods for nonlinear equations, Intern. J. Computer Math., 9 (1981), 353-361.

25. Neta, B., A new iterative method for the solution of systems of nonlinear equations, Proc. Approx. Th. Applics (z. Ziegler, ed.), Academic Press, NY, 1981, 249-263.

26. Neta, B., A note on Neta's sixth-order family of methods for solving equations, Intern. J. Computer Math., to appear Vol. 14.

27. Neta, B., Victory, H.D., A higher order method for determining nonisolated solutions of a system of nonlinear equations, to appear.

28. Ortega, J., Rheinboldt, W.C., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, NY, 1970.

29. Ostrowski, A.M., Solution of Equations and Systems of Equations, 3rd ed. Academic Press, New York-London, 1973.

30. Popovski, D.B., A family of one-point iteration formulae for finding roots, Inter. J. Computer Math., 8 (1980), 85-88.

31. Popovski, D.B., An extension of Chebyshev's iteration, Informatica, 4 (1980), 26-28.

32. Popovski, D.B., A note on King's fifth order family of methods for solving equations, BIT, 21 (1981), 129-130.

33. Popovski, D.B., A note on Neta's family of sixth-order methods for solving equations, Inter. J. Computer Math., 10 (1981), 91-93.

34. Popovski, D.B., Sixth order methods for solving equations. J. Appl. Math. Phys. (ZAMP), 33 (1982), 434-438.

35. Rall, L., Convergence of the Newton process to multiple solutions, Numer. Math. 9 (1966), 23-37.

36. Reddien, G.W., On Newton's method for singular problems, SIAM J. Numer. Anal. 15 (1978), 993-996.

37. Reddien, G.W., Newton's method and high order singularities, Computer and Math. with Applics. 5 (1979), 79-86.

38. Rheinboldt, W.C., Methods for Solving Systems of Nonlinear Equations, SIAM, Philadelphia, 1974.

39. Schröder, J., Über das Newtonsche verfahren, Arch. Rational Mech. Anal., 1 (1957), 154-180.

40. Steffensen, J.F., Remarks on iteration, Skand. Aktuar. Tidskr., 16 (1934), 64

41. Traub, J.F., Iterative Methods for the Solution of Equations. Prentice-Hall, New York, 1964.

42. Victory, H.D., Neta, B., A higher order method for multiple zeros of nonlinear functions. Intern. J. Computer Math., 12 (1982), 329-335.

43. Voigt, R.G., Rates of convergence for a class of iterative procedures, SIAM J. Numer. Anal. 8 (1971), 127-134.

44. Voigt, R.G., Orders of convergence for iterative procedures, SIAM J. Numer. Anal. 8 (1971), 222-243.

45. Weber, H., Werner, W., On the accurate determination of nonisolated solutions of nonlinear equations, Computing 26 (1981), 315-326.

46. Werner, W., Über ein Verfahren der Ordnung $1 + \sqrt{2}$ zur Nullstellenbestimmung, Numer. Math. 32 (1979), 333-342.

47. Werner, W., Some improvements of classical iterative methods for the solution of nonlinear equations, Proc. Numer. Solution of Nonlinear Equations, Bremen, 1980, E.L. Allgower, K. Glashoff and H.O. Peitgen eds.

Bounds on the dimension of spaces of multivariate piecewise polynomials

by

Larry L. Schumaker*

CAT #25 October,1982

§1. Introduction.

Spaces of piecewise polynomials defined over a partition of a planar set are of considerable interest in approximation theory. In addition to their usefulness in a variety of data fitting problems, they also play a central role in the finite element method. Clearly, they are a natural generalization of the classical one-dimensional polynomial spline functions.

Despite their obvious importance, until recently there has been relatively little work on general spaces of piecewise polynomials in two variables. In the last few years, however, the literature has grown considerably -- see [1 -25] and references therein.

Some years ago in [19], I gave a lower bound on the dimension of spaces of piecewise polynomials defined on a triangulation. The purpose of this paper is to present both lower and upper bounds for general rectilinear partitions. The plan of the paper is as follows. In the remainder of this section we introduce the spline spaces of interest and

---

* Supported in part by NASA grant 4764-2

establish some notation. In Sections 2 and 3 we establish our upper and lower bounds, respectively. Section 4 of the paper contains a variety of applications to special partitions. We conclude the paper with remarks.

Suppose Ω is a closed subset of R^2, and suppose that $\Delta = \{\Omega_i\}$ is a collection of open subsets such that

$$1) \qquad \overline{\Omega} = \bigcup_1^n \Omega_i$$

$$2) \qquad \Omega_i \cap \Omega_j = \emptyset \quad , \text{ all } i,j = 1,2,\ldots,n.$$

We call Δ a <u>partition</u> of Ω. If each Ω_i is a polygon, then we call Δ a <u>rectilinear partition.</u> If each Ω_i is a triangle and if no vertex of any triangle lies in the middle of an edge of another triangle, then we call Δ a <u>triangulation</u> of Ω.

Given a positive integer d we define the space of polynomials of order d (in two variables) by

$$P_d = \{p(x,y) = \sum_{i=0}^{d} \sum_{j=0}^{d-i} a_{ij} x^i y^j \ , \ a_{ij} \in R \}.$$

<u>Definition 1.1.</u> Let $0 < r < d$, and set

$$(1.1) \qquad S_d^r(\Delta) = \{s \in C^r(\Omega): s\big|_{\Omega_i} \in P_d \ , \ i = 1,\ldots,n\}$$

We call S the space of <u>polynomial splines of order d and smoothness r</u> associated with the partition Δ.

It is clear that S is a linear space. In this paper we are interested in computing its dimension. It turns out that it is not possible to give a general formula -- there are some cases where the dimension depends on the exact geometry of the partition -- see [17,19]. In general, we must be satisfied with upper and lower bounds for it.

§2. An upper bound on dimension.

In order to state our main result, we need some additional notation. Throughout the remainder of this section we shall suppose that Δ is a rectilinear partition of a set Ω. Given such a partition, we call the straight line segments making up the partition the edges of the partition, and refer to the points where these edges join together as the vertices of the partition. We denote the number of edges and the number of vertices in the interior of Ω by E and V, respectively.

Associated with the integers d and r, we define

$$\alpha = (d+1)(d+2)/2 \ , \quad \beta = (d-r)(d-r+1)/2 \tag{2.1}$$

and

$$\gamma = [(d+1)(d+2)-(r+1)(r+2)]/2 \ . \tag{2.2}$$

We are now ready for the main result of the paper.

Theorem 2.1. Suppose that the vertices of the partition Δ are numbered in such a way that each pair of consecutive vertices in the list are corners of a common subset in Δ. For each $i = 1,2,\ldots,V$, let

$\tilde{e}_i$ = number of edges with different slopes attached to the i-th vertex but not attached to any of the first i-1 vertices in the list, (2.3)

and let

$$\tilde{\sigma}_i = \sum_{j=1}^{d-r} (r+j+1 - j\cdot\tilde{e}_i)_+ \ . \tag{2.4}$$

Then

$$\dim S_d^r(\Delta) \le \alpha + \beta E - \gamma V + \sum_{i=1}^{V} \tilde{\sigma}_i \ . \tag{2.5}$$

Proof: Let N be the number on the right-hand side of (2.5). By an elementary lemma of linear algebra (cf. Lemma 3.3 of [19]), it suffices to construct linear functionals $\lambda_1,\ldots,\lambda_N$ such that

(2.6) if $s \in S$ and $\lambda_i s = 0$, $i = 1,\ldots,N$, then $s \equiv 0$.

Suppose the vertices of the partition are $\xi_1,\ldots,\xi_V$, and let Ω^0 be a set in Δ with a corner at ξ_1 (cf. Figure 1). It is well-known (cf. [16]) that we can find a set Λ^0 of α point functionals in Ω^0 which anihilate P_d. Let E_1 be the number of edges attached to the vertex ξ_1. We claim that we can find an additional $\beta E_1 - \gamma + \tilde{\sigma}_1$ functionals to obtain a set Λ^1 which anihilates any function in $s|_{\Omega 1}$, where

$$\Omega^1 = \bigcup \{\Omega_i : \Omega_i \text{ has a vertex at } \xi_1 \} .$$

To show this, we may suppose that ξ_1 is at the origin, and that the figure is rotated so that none of the edges lie on the x or y axes. Suppose we number the edges counter clockwise, starting from Ω^0. Each of these edges is described by an angle θ_j or equivalently by an equation $y+\alpha_j x = 0$, $j = 1,\ldots,E_1$.

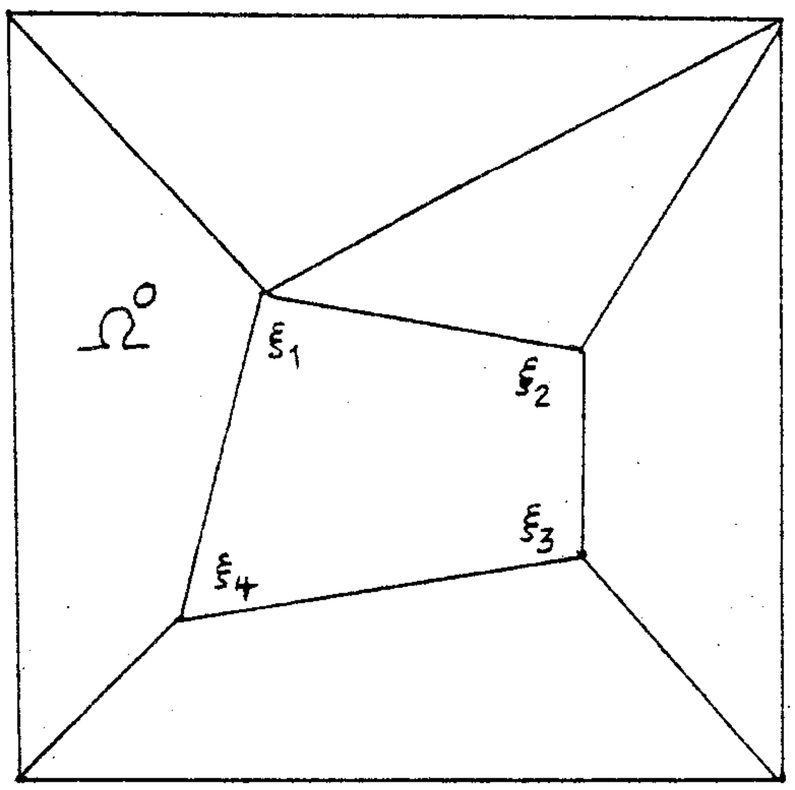

Figure 1. A rectilinear partition.

Now suppose $s \equiv 0$ on Ω^0. Then after crossing the first edge, s must have the form

$$(2.7)\quad s(x,y) = \sum_{j=1}^{d-r} \sum_{k=1}^{j} c_{j,k}\ \phi^1_{j,k}(x,y),$$

where in general

$$\phi^i_{j,k}(x,y) = y^{j-k}(y+\alpha_i x)^{r+k}.$$

Define

$$\phi^i_{j,k}(x,y)_+ = \begin{cases} \phi^i_{j,k}(x,y), & \text{if } \arctan(y/x) \geq \theta_i \\ 0, & \text{otherwise.} \end{cases}$$

Then it follows that the set of functions in $S|_{\Omega 1}$ which vanish on Ω^0 are precisely the functions of the form

$$s(x,y) = \sum_{\nu=1}^{n} \sum_{j=1}^{d-r} \sum_{k=1}^{j} c_{\nu jk}\phi^{\nu}_{jk}(x,y)_+, \qquad (n = E_1)$$

where the coefficients satisfy the equations

$$\sum_{\nu=1}^{n} \sum_{j=1}^{d-r} \sum_{k=1}^{j} c_{\nu jk}\phi^{\nu}_{jk}(x,y) = 0 \quad \text{for all } x,y \in \Omega^0 .$$

By equating the coefficients of the various powers of $x^\nu y^\mu$ to zero, we can rewrite this as a homogeneous system of linear equations

$$(2.8)\qquad Ac = 0,$$

where

$$c = (c_1,\ldots,c_{d-r})^T$$

$$c_j = (c_{1j1},\ldots,c_{1jj},\ldots,c_{nj1},\ldots,c_{njj})^T,$$

and

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_{d-r} \end{bmatrix},$$

where for each $j = 1,\ldots,d-r$, A_j is an $r+j+1$ by $n \cdot j$ matrix of the form

$$A_j = [A_{j1},\ldots,A_{jn}]$$

with A_{ji} given by

$$\begin{bmatrix} 1 & & & \\ \binom{r+j}{1}\alpha_i & 1 & & \\ \binom{r+j}{2}\alpha_i^2 & \binom{r+j-1}{1}\alpha_i & 1 & \\ \cdots & & & \\ \binom{r+j}{r+j}\alpha_i^{r+j} & \binom{r+j-1}{r+j-1}\alpha_i^{r+j-1} & \cdots & \binom{r+1}{r+1}\alpha_i^{r+1} \end{bmatrix},$$

Clearly (2.7) is a system of γ equations in $n\beta$ unknowns. It is shown in [19] that the rank of A is $\gamma - \tilde{\sigma}_1$. We conclude that we can add $n\beta - \gamma + \tilde{\sigma}_1$ equations to force the coefficients to be zero. Since $n = E_1$, this is equivalent to adding $\beta E_1 - \gamma + \tilde{\sigma}_1$ linear functionals to Λ^0 to get Λ^1 .

We now continue this process one vertex at a time. In particular, if E_i denotes the number of edges attached to the vertex ξ_i (but not to any of the vertices $\xi_1,\ldots,\xi_{i-1}$), then we can add a set of $\beta E_i - \gamma + \tilde{\sigma}_i$ functionals to Λ^{i-1} to get a set Λ^i which anihilates splines on

$$\Omega^i = \Omega^{i-1} \cup \{ \Omega_j : \Omega_j \text{ has a corner at } \xi_i \} .$$

After proceeding through all vertices and adding β functionals associated with each remaining uncounted edge, we end up with a set of N linear functionals which anihilates all of S. This completes the proof. ■

For convenience, we give values of α,β,and γ in Table 1 for several choices of d and r.

d	r	α	β	γ
2	1	6	1	3
3	1	10	3	7
4	1	15	6	12
5	1	21	10	18
3	2	10	1	4
4	2	15	3	9
5	2	21	6	15
4	3	15	1	5
5	3	21	3	11

Table 1. The coefficients (2.1)-(2.2) for some choices of d and r.

It is clear that the upper bound given in Theorem 2.1 is numerically computable. The following example shows that its value depends on the ordering of the vertices.

Example 2.2. Let Ω and Δ be as shown in Figure 2, and let $d = 2$ and $r = 1$. Compute an upper bound for the associated spline space.

Discussion: Here $\alpha = 6$, $\beta = 1$, and $\gamma = 3$. If we order the vertices so that the lower one comes first, then we have $\tilde{\sigma}_1 = (3-\tilde{e}_1)_+ = 0$ and $\tilde{\sigma}_2 = (3-\tilde{e}_2)_+ = 1$,and hence

$$\dim S_2^1(\Delta) \leq 6 + 7 - 2\cdot 3 + 1 = 8.$$

On the other hand, if we order the vertices so that the upper one comes first, then $\tilde{\sigma}_1 = \tilde{\sigma}_2 = 0$, and now

$$\dim S_2^1(\Delta) \leq 6 + 7 - 2\cdot 3 = 7. \quad \blacksquare$$

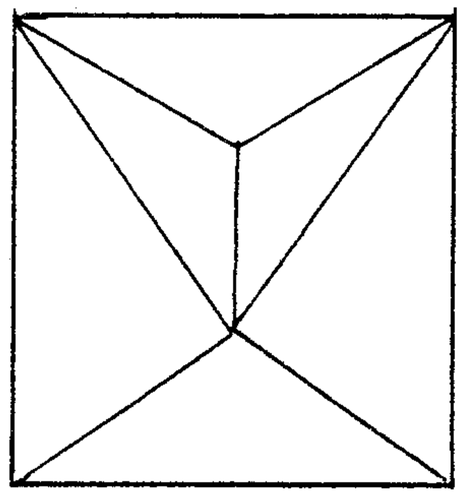

Figure 2. The partition for Example 2.2.

§3. A lower bound on dimension.

In order to be able to use the upper bound of Theorem 2.1 to determine the exact dimension of a space of splines, we need to have a lower bound to combine it with. We begin by presenting a lower bound which applies to arbitrary rectilinear partitions.

Theorem 3.1. Let Δ be a rectilinear partition of a set Ω, and let α,β and γ be as in (2.1)-(2.2). Given any ordering of the vertices, let

(3.1) e_i = number of edges with different slopes attached to the i-th vertex,

$$(3.2)\quad \sigma_i = \sum_{j=1}^{d-r} (r+j+1 - j\cdot e_i)_+ \quad , \quad i = 1,\ldots,V.$$

Then

$$(3.3)\quad \dim S_d^r(\Delta) \geq \alpha + \beta E - \gamma V + \sum_{i=1}^{V} \sigma_i \quad .$$

Proof: The proof follows along the same lines as the proof of Theorem 3.1 of [19]. In particular, if Δ is a partition of a set Ω with only one interior vertex, then the argument proceeds exactly as before. To get the result for a general partition, we use a merging procedure. Indeed, given

any vertex on the boundary of Ω connected to an interior vertex ξ^* by an edge, we may remove one polygon having these vertices to get a new set Ω_1 with one less interior vertex (cf. Figure 3). Assuming the result for partitions with V-1 interior vertices, we can then merge this space with a spline space over the cell Ω_2 with interior vertex ξ^* to get the result -- cf. the argument in [19]. ■

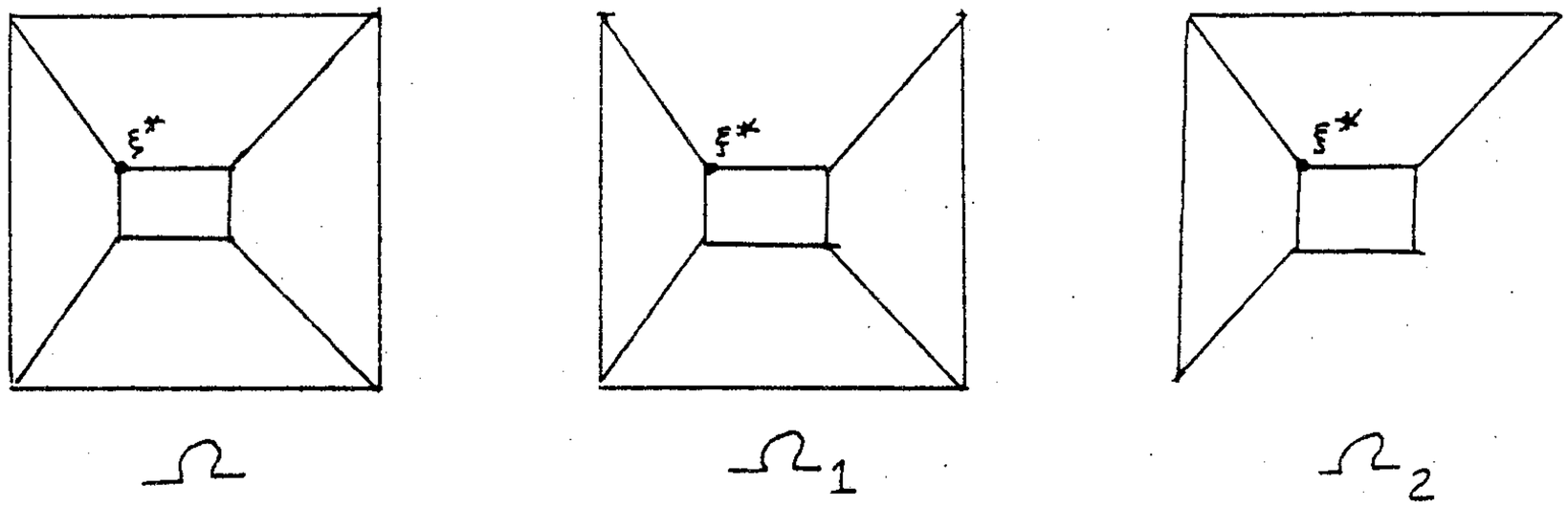

Figure 3. Merging cells of a partition.

We note that the lower bound in (3.3) has exactly the same form as the upper bound in (2.5), the only difference being that here σ_i replaces $\tilde{\sigma}_i$. By the definitions (2.4) and (3.2), it is clear that for each $i = 1,\ldots,n$, $\sigma_i \leq \tilde{\sigma}_i$, and thus the upper bound is greater than or equal to the lower bound. We have the following immediate corollary.

Corollary 3.2. Suppose that for some ordering of the vertices of a triangulation Δ we have

$$(3.4) \qquad \sigma_i = \tilde{\sigma}_i , \quad i = 1,\ldots,V.$$

Then the expressions in (2.5) and (3.3) agree and give the dimension of the spline space (1.1).

The following example (cf. [17,19]) shows that even for relatively simple triangulations, it may happen that our lower and upper bounds do not agree.

Example 3.3 Let Δ be the triangulation show in Figure 4, and let $d = 2$ and $r = 1$. Compute the dimension of the corresponding spline space.

Discussion: It is easy to see that $\sigma_1 = \sigma_2 = \sigma_3 = 0$ while no matter how we order the vertices, we will always have $\tilde{\sigma}_3 = 1$. It follows that

$$6 \leq \dim S_2^1(\Delta) \leq 7 \quad .$$

Indeed, it is known (cf. [17,19]) that the exact dimension of this spline space depends on the location of the vertices. If the figure is symmetric, the dimension is 7; otherwise it is 6. ■

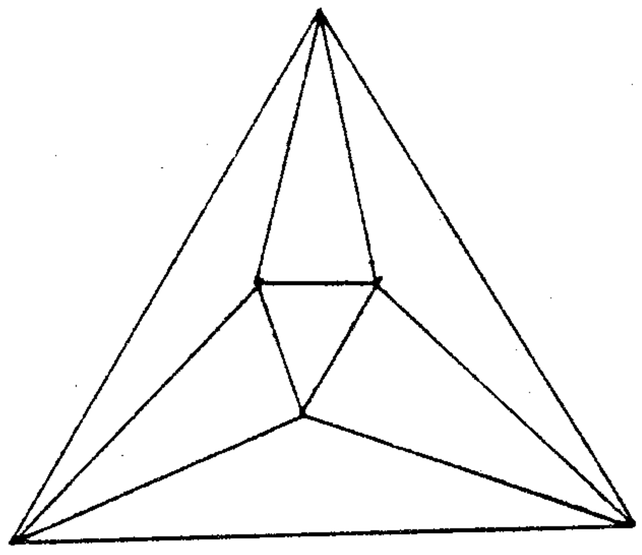

Figure 4. The partition in Example 3.4.

We now give an example of a rectangular partition where the upper and lower bounds do agree. Other examples can be found in Section 5.

Example 3.4 Let Δ be the rectilinear partition shown in Figure 5. Let $d = 2$ and $r = 1$. Compute the dimension of the associated spline space.
Discussion: It is easily seen that if we number the vertices as shown in Figure 5, then $\tilde{\sigma}_i = (3-\tilde{e}_i)_+ = 0$, $i = 1,\ldots,4$. It follows that

$$\dim S_2^1(\Delta) = 6 + 17 - 4\cdot 3 = 11. \quad \blacksquare$$

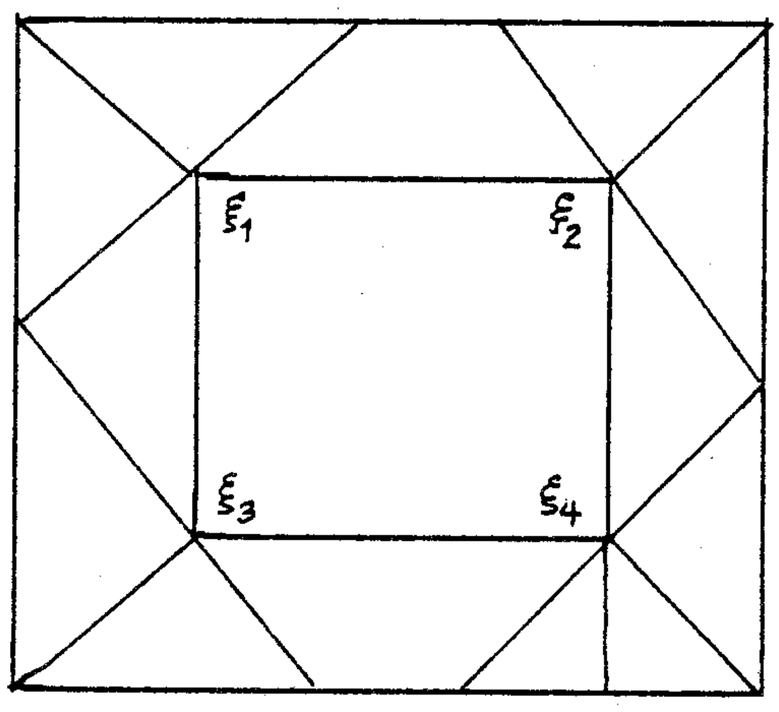

Figure 5 The partition in Example 3.4.

§4. Examples and applications.

In this section we apply our upper and lower bounds to several cases of interest. We begin with a special partition of a rectangle. Let

$$\Omega = [a,b] \times [\tilde{a},\tilde{b}]$$
$$a = x_0 < x_1 < \ldots < x_k < x_{k+1} = b \tag{4.1}$$
$$\tilde{a} = \tilde{x}_0 < \tilde{x}_1 < \ldots < \tilde{x}_{\tilde{k}} < \tilde{x}_{\tilde{k}+1} = \tilde{b}$$

If Δ is the partition of Ω which is obtained by drawing grid lines at the points $x_1,\ldots,x_k$ and $\tilde{x}_1,\ldots,\tilde{x}_{\tilde{k}}$ along with the upward sloping diagonals (cf. Figure 6), then we call Δ a <u>type one partition.</u>

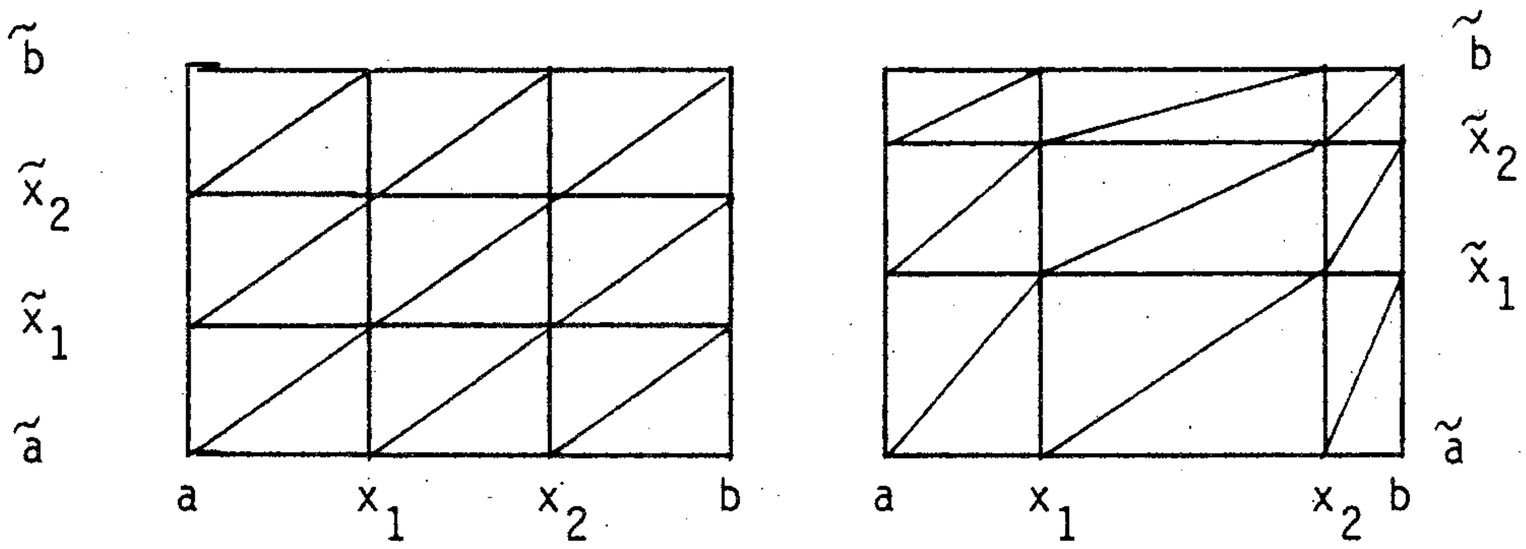

Figure 6 Type-1 partitions.

<u>Theorem 4.1.</u> Let Δ be an equally spaced type-1 partition of a rectangle Ω. Then for all $0 \le r < d$,

$$\dim S_d^r(\Delta) = k\tilde{k}(d^2 - 3rd + 2r^2 + \sigma) + (k+\tilde{k})(d^2 - 2rd + d - r + r^2) + (2d^2 + 4d - 2rd - r + r^2 + 2)/2 , \tag{4.2}$$

where

$$\sigma = \begin{cases} r^2/4 & , \text{if } r \text{ is even and } 3r+1 > 2d \\ (r^2-1)/4 & , \text{if } r \text{ is odd and } 3r+1 > 2d \\ (d-r)(2r-d) & , \text{otherwise.} \end{cases} \tag{4.3}$$

<u>Proof:</u> We easily check that $V = k\tilde{k}$ and $E = 3k\tilde{k} + 2(k+\tilde{k}) + 1$. If we put the vertices of the partition in lexicographical order, then we note that $e_i = \tilde{e}_i = 3$ for all i, and thus

$$\sigma_i = \tilde{\sigma}_i = \sigma := \sum_{j=1}^{d-r} (r+1-2\cdot j)_+$$

for all $i = 1,\ldots,V$. It follows that Corollary 3.2 can be applied, and after some algebra, we obtain (4.2). ■

Our next theorem deals with arbitrary type-1 partitions.

Theorem 4.2. Let Δ be a general type-1 partition. Then for all $1 < d$,

$$\dim S_d^1(\Delta) = k\tilde{k}(d^2 - 3d + 2) + (k+\tilde{k})(d^2-d) + (d^2+d+1) \quad . \tag{4.4}$$

Proof: If we put the vertices in lexicographical order, then it is easy to see that $\tilde{e}_i$ is always at least 3, and thus $\tilde{\sigma}_i = 0$ for $i = 1,\ldots,V$. Now applying Corollary 3.2, we obtain (4.4). ■

Theorem 4.2 gives the dimension of C^1 spline spaces on general type-1 partitions. The following example shows that for $r > 1$, our upper and lower bounds do not agree, and in fact the actual dimension can be equal to the upper bound.

Example 4.3. Let Δ be the unequally spaced type-1 partition shown in Figure 6 with $k = \tilde{k} = 2$. Let $d = 3$ and $r = 2$.

Discussion: If we put the vertices in lexicographical order, then we see that $\sigma_i = 0$ and $\tilde{\sigma}_i = \delta_{i3}$, $i = 1,\ldots,4$. It follows that

$$15 \leq \dim S_3^2(\Delta) \leq 16 \quad .$$

The actual dimension of this space is 16. Indeed, in addition to the 10 linearly independent polynomials in S, the following six splines also belong to the space:

$$(x-x_1)_+^3 \quad , \quad (x-x_2)_+^3 \quad , \quad (y-\tilde{x}_1)_+^3 \quad , \quad (y-\tilde{x}_2)_+^3 \quad ,$$

$$(y - \tilde{x}_2 - (x-x_0)(\tilde{x}_3 - \tilde{x}_2)/(x_1-x_0))_+^3 \quad , \quad (y - (x-x_2)(\tilde{x}_1-\tilde{x}_0)/(x_3-x_2))_+^3$$

Theorem 4.1 asserts that this space has dimension 19 in the case where Δ is an equally spaced type-1 partition. ■

We turn now to another special partition of a rectangle. Let Ω and grid points $x_1,\ldots,x_k$ and $\tilde{x}_1,\ldots,\tilde{x}_{\tilde{k}}$ be given as in (4.1). If Δ is a partition of Ω which is obtained by drawing all the grid lines plus both diagonals in each subrectangle, we call Δ a <u>type-2 partition</u> of Ω. Typical type-2 partitions are shown in Figure 7.

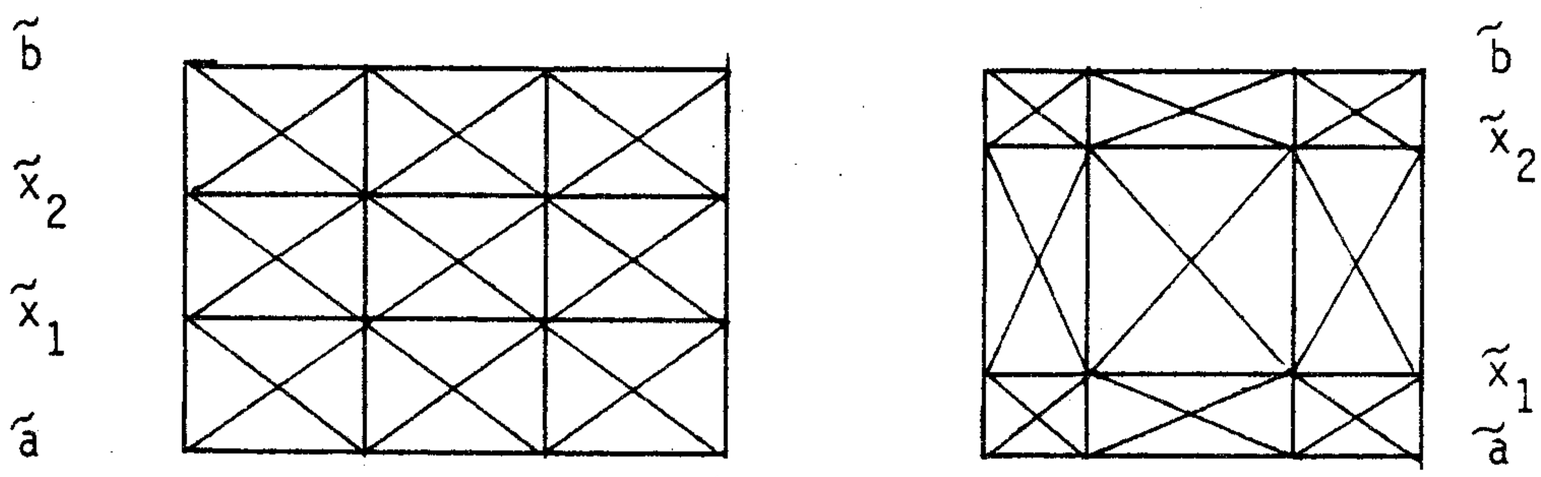

Figure 7 Type-2 partitions.

<u>Theorem 4.4.</u> Let Δ be an equally spaced type-2 partition of a rectangle Ω. Then for all $0 \le r < d$,

$$\begin{aligned}(4.5)\qquad \dim S_d^r(\Delta) = {} & k\tilde{k}(2d^2 - 6rd + 4r^2 + \sigma_g + \sigma_c) \\ & + (k+\tilde{k})(2d^2 - 5rd + d - r + 3r^2 + \sigma_c) \\ & + (4d^2 + 4d - 8rd - r + 5r^2 + 2 + 2\sigma_c)/2 ,\end{aligned}$$

where

$$(4.6)\qquad \sigma_g = \sum_{j=1}^{d-r} (r + 1 - 3\cdot j)_+ \quad , \quad \sigma_c = \sum_{j=1}^{d-r} (r+1-j)_+ .$$

<u>Proof:</u> Here there are $k\tilde{k}$ vertices at the corners of the grid and an additional $k\tilde{k} + (k+\tilde{k}) + 1$ vertices where the diagonals cross. The number of edges is given by $E = 6k\tilde{k} + 5(k+\tilde{k}) + 4$. If we put the grid vertices in lexicographical order, followed by the cross vertices, then we note that $\sigma_i = \tilde{\sigma}_i = \sigma_g$ for $i = 1,\ldots,k\tilde{k}$. For the remaining points we have $\sigma_i = \tilde{\sigma}_i = \sigma_c$, $i = k\tilde{k} + 1,\ldots,V$. Now Corollary 3.2 applies, and after some algebra, we obtain (4.5). ■

The above theorem deals with equally-space type-2 partitions. In the unequally spaced case, we have the following result.

Theorem 4.5. Let Δ be an arbitrary type-2 partition of a rectangle Ω. Then for all $r < d$ with $r = 0,1,2,$

(4.7) $\dim S_d^r(\Delta)$ is given by the formula (3.3).

Proof: If we order the vertices as in the proof of Theorem 4.4, then since $\tilde{e}_i \geq 4$, we see that $\tilde{\sigma}_i = 0$, $i = 1,\ldots,k\tilde{k}$. On the other hand, since $e_i = \tilde{e}_i = 2$ and thus $\sigma_i = \tilde{\sigma}_i$ for $i = k\tilde{k}+1,\ldots,V$, Corollary 3.2 applies to establish the result. ■

We close this section with a result on a more general kind of partition. Let Ω be the closure of an arbitrary domain, and suppose that a partition Δ is a simple cross cut partition (cf. [2]) obtained by drawing L lines across Ω. (Being simple requires that exactly two lines meet at each vertex in Ω -- see Figure 8).

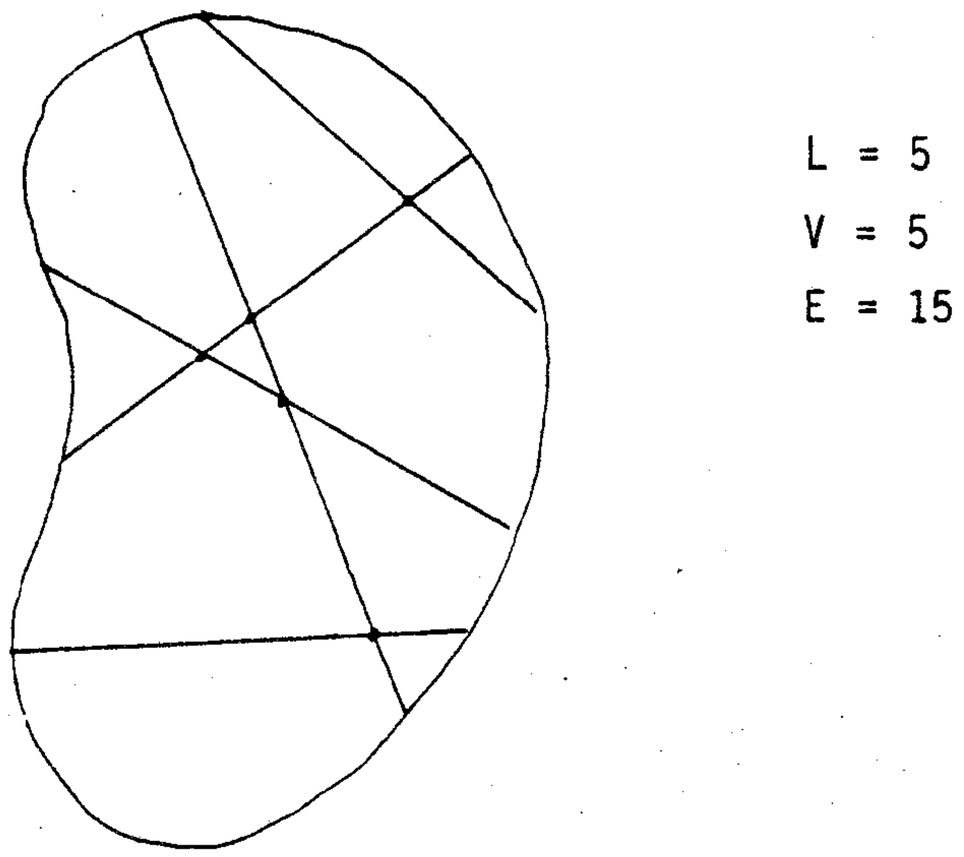

Figure 8 A simple cross-cut partition.

Theorem 4.6. Let Δ be a simple cross-cut partition of a set Ω, and let α and β be as in (2.1). Then for all $0 \leq r < d < 2r+1$,

$$(4.8) \quad \dim S_d^r(\Delta) = \alpha + \beta E - V(d^2 + 3d - 2r^2 - 4r)/2,$$

while if $2r+1 \leq d$, then

$$(4.9) \quad \dim S_d^r(\Delta) = \alpha + \beta(E - 2V) \quad .$$

Proof: Here $e_i = \tilde{e}_i = 2$ for $i = 1,\ldots,V$, and so

$$\sigma_i = \tilde{\sigma}_i = \sum_{j=1}^{d-r} (r+1-j)_+ = \begin{cases} r(r+1)/2 & , \text{ if } 2r+1 < d \\ -3r^2-r+d+4dr-d^2 & , \text{ if } r < d \leq 2r. \end{cases}$$

The result now follows from Corollary 3.2. ■

The result (4.9) agrees with the formula obtained in Theorem 5.1 of [2] when we take note of the fact (cf. Lemma 5.1 of [2]) that $L = E - 2V$.

§5. Remarks.

1. In this paper we have confined our attention to rectilinear partitions since it is very difficult to see what the connection would be between polynomials in adjoining regions separated by a curved boundary. The boundary of Ω itself can, of course, be curved.

2. The idea of obtaining an upper bound on dimension by placing linear functionals was used already in several earlier papers -- see eg. [17-19]. Despite using them in a similar way on some special cases in [19], I did not see the general result at the time, however.

3. Generally I have followed the notation of [19] throughout this paper. One notable change, however, is that here I am using d for the degree of the polynomials, while in [19] I used m for the order. The two are connected by $m = d+1$.

4. Corollary 3.2 is easiest to apply when the number of edges at each vertex is relatively large compared with the smoothness order r. Unfortunately, it is easy to construct a variety of examples similar to Example 3.3 where the upper and lower bounds do not agree. On the other hand, there are also many examples where they do agree and provide a dimension statement in situations where no other presently available methods apply. Example 3.4 is a case which does not seem to fit any available theory, not even the quasi-cross-cut theory of [6]. It is of interest to note that in this example the correct dimension equals the upper bound rather than the lower one.

5. I had hoped that the upper bound presented here would shed some light on why high degree splines with low smoothness do not seem to be subject to the difficulty inherent in Example 3.3. (It is known [17] that for all $d \geq 5$, the C^1 splines on the partition in Figure 4 have dimension given by the lower bound. The cases d = 3,4 remain unclarified).

6. Type-1 and Type-2 partitions have been considered in a variety of papers -- see e.g. [3-10,19]. Dimension statements for the equally spaced case and for several values of d and r can be found in [19]. Slightly different looking (but equivalent) formulae were found for general d and r in [6]. The results for unequally spaced partitions are new.

7. In comparing the upper and lower bounds given here for a variety of quasi-cross-cut partitions, I found them to agree with each other and with the dimensionality formulae given in [6]. Hence, I conjecture that this holds for general quasi-cross-cut partitions. The problem in constructing a proof is that there does not seem to be a simple relation between vertices, edges, and lines in such a partition.

8. Recently there have appeared a number of results on the dimension of spaces of splines which satisfy boundary conditions -- see [8-10]. The tools presented here can also be used on these kinds of spline spaces.

9. After identifying the dimension of a space of splines, the next important question is to construct a basis, and if possible a local basis, for the space. Considerable work has been done on this problem for special partitions -- see e.g. [1-10] and references therein.

References

1. Chui, C.K. and R.H. Wang, Bases of bivariate spline spaces with cross-cut grid partitions, J. Math. Research and Exposition, to appear.

2. Chui, C.K. and R.H. Wang, On smooth multivariate spline functions, CAT Rpt. # 3, 1981.

3. Chui, C.K. and R.H. Wang, Bivariate cubic B-splines relative to cross-cut triangulations, CAT Rpt. #4, 1981.

4. Chui, C.K. and R.H. Wang, Multivariate B-splines on triangulated rectangles, CAT Rpt. #6, 1981.

5. Chui, C.K. and R.H. Wang, On a bivariate B-spline basis, CAT #7, 1981.

6. Chui, C.K. and R.H. Wang, Multivariate spline spaces, CAT #9, 1981.

7. Chui, C.K. and R.H. Wang, Spaces of bivariate cubic and quartic splines on type-1 triangulations, CAT #20, 1981.

8. Chui, C.K., and L.L. Schumaker, On spaces of piecewise polynomials with boundary conditions. I. Rectangles, CAT #17, 1982.

9. Chui, C.K., L.L. Schumaker, and R.H. Wang, On spaces of piecewise polynomials with boundary conditions. II. Type-1 triangulations, CAT #18, 1982.

10. Chui, C.K., L.L. Schumaker, and R.H. Wang, On spaces of piecewise polynomials with boundary conditions. III. Type-2 triangulations, CAT #19, 1982.

11. Fredrickson, P., Triangular spline interpolation, Rpt. #670, Whitehead Univ., 1970.

12. Fredrickson, P., Generalized triangular splines, Rpt. #7-71, Lakehead Univ., 1971.

13. Heindl, G., Uber verallgemeinerte Stammfunktionen und LC-Funktionen in R^n, dissertation, Tech. Univ. Munich, 1968.

14. Heindl, G., Spline-Funktionen mehrerer Veranderlicher. I., Bayerische Akad. 6(1970), 49-63.

15. Heindl, G., Interpolation and approximation by piecewise quadratic C^1 functions of two variables, in Multivariate Approximation Theory, W. Schempp and K. Zeller, eds., Birkhauser, Basel, 1979, 146-161.

16. Morgan, J. and R. Scott, A nodal basis for C^1 piecewise polynomials of degree $n \geq 5$, Math. Comp. 29(1975), 736-740.

17. Morgan, J. and R. Scott, The dimension of piecewise polynomials, manuscript, 1977.

18. Powell, M.J.D., Piecewise quadratic surface fitting for contour plotting, in Software for Numerical Analysis, D.J. Evans, ed., Academic Press, N.Y., 1974, 253-271.

19. Schumaker, L.L., On the dimension of spaces of piecewise polynomials in two variables, in Multivariate Approximation Theory, W. Schempp and K. Zeller, eds., Birkhauser, Basel, 1979, 396-412.

20. Schumaker, L.L., Spline Functions: Basic Theory, Wiley-Interscience, N.Y., 1981.

21. Strang, G., The dimension of piecewise polynomials, and one-sided approximation, in Lecture Notes 365, Springer-Verlag, N.Y., 1974, 144-152.

22. Wang, R.H., The structural characterization and interpolation for multivariate splines, Acta Math. Sinica 18(1975), 91-106.

23. Wang, R.H., On the analysis of multivariate splines in the case of arbitrary partition, Sci. Sinica (Math. I), 1979, 215-226.

24. Wang, R.H., On the analysis of multivariate splines in the case of arbitrary partition II, Num. Math. of China 2(1980), 78-81.

25. Zwart, P., Multi-variate splines with non-degenerate partitions, SIAM J. Numer. Anal. 10(1973), 665-673.

ABSTRACT

On the Simplification of Generalized Conjugate Gradient Methods for Nonsymetric Systems*

by

Kang C. Jea
Fu Jen University
Tapei, Taiwan

and

David M. Young
The University of Texas at Austin
Austin, Texas, U.S.A.

This is the summary of a talk given by David M. Young at the Thirteenth Annual Mathematics Conference at The University of Southwestern Louisiana, Lafayette, Louisiana, October 23, 1982. Further details will be found in the paper by Kang C. Jea and David M. Young to appear in the journal Linear Algebra and its Applications.

---

* The work was supported in part by the Natural Science Foundation under Grant MCS-7919829, and by the Department of Energy under Grant DE-AS05-81ER10954 with The University of Texas at Austin.

The paper is concerned with generalized conjugate gradient methods for solving the linear system $Au = b$ where A is a given $N \times N$ matrix, b is a given $N \times 1$ vector, and the $N \times 1$ vector u is to be determined. The methods used can be regarded as generalizations of the conjugate gradient method (CG method) developed by Hestenes and Stiefel [6] which is applicable in the symmetrizable case where HA is symmetric and positive definite (SPD) for some SPD matrix H. The convergence of the CG method compares favorably with that of a number of other iterative methods, and, moreover, at each step the determination of a new iteration vector only involves the use of information from, at most, two preceding iterations. Young and Jea [11] and Jea [7] considered a method called the idealized generalized conjugate gradient method (IGCG method) for handling nonsymmetrizable linear systems. The IGCG method is defined by choosing a nonsingular auxiliary matrix Z and requiring that $u^{(n)} - u^{(0)} \in K_n(r^{(0)})$ where $K_n(r^{(0)})$ is the Krylov space spanned by $r^{(0)}, Ar^{(0)}, \ldots, A^{n-1}r^{(0)}$. Here $r^{(0)}$ is the residual vector $b - Au^{(0)}$. One also requires that $(Zr^{(n)}, v) = 0$ for all $v \in K^n(r^{(0)})$.

(If ZA is SPD, one can require instead that $\|u^{(n)} - A^{-1}b\| \leq \|v - A^{-1}b\|$

for all v such that $v - u^{(0)} \in K_n(r^{(0)})$. Here the norm is defined by $\|w\| = (w, ZAw)^{\frac{1}{2}}$.

Young and Jea [11] gave three forms of the IGCG method. The first form, called ORTHODIR, has the formulas

$$u^{(n+1)} = u^{(n)} + \hat{\lambda}_n q_n \tag{1}$$

where the direction vectors, $q^{(0)}$, $q^{(1)}$, ... , have the form

$$q^{(n)} = Aq^{(n-1)} + \beta_{n,n-1}q^{(n-1)} + \ldots + \beta_{n,0}q^{(0)}, \quad (q^{(0)} = r^{(0)}) \tag{2}$$

Here the coefficients $\beta_{n,n-1}, \ldots, \beta_{n,0}$ are chosen so that $(ZAq^{(i)}, q^{(j)}) = 0$ for $i > j$. If, for some integer $s \geq 1$, one requires that $\beta_{n,i} > 0$ for $i < n-s$ we have a truncated version of ORTHODIR, called ORTHODIR(s). In the symmetrizable case, if Z and ZA are SPD, then ORTHODIR reduces to ORTHODIR(2). The second form of the IGCG method, called ORTHOMIN has the formulas

$$u^{(n+1)} = u^{(n)} + \lambda_n p^{(n)} \tag{3}$$

where the direction vectors $p^{(0)}$, $p^{(1)}$, ... , have the form

$$p^{(n)} = r^{(n)} + \alpha_{n,n-1}p^{(n-1)} + \ldots + \alpha_{n,0}p^{(0)}, \quad (p^{(0)} = r^{(0)}) \tag{4}$$

Here $r^{(n)} = b - Au^{(n)}$ and the coefficients $\alpha_{n,n-1}, \ldots, \alpha_{n,0}$ are chosen so that $(ZAp^{(i)}, p^{(j)}) = 0$ for $i > j$. If for some integer $s \geq 1$, we require that $\alpha_{n,i} = 0$ for $i < n-s$ we have the truncated scheme, called ORTHOMIN(s). In the symmetrizable case, if Z and ZA are SPD, then ORTHOMIN reduces to ORTHOMIN(1). ORTHOMIN(1) is the classical form of the CG method given by Hestenes and Steifel [6]. The third form of the IGCG method, called ORTHORES has the formulas

$$u^{(n+1)} = \lambda_n r^{(n)} + f_{n+1,n} u^{(n)} + f_{n+1,n-1} u^{(n-1)} + \ldots + f_{n+1,0} u^{(0)} \tag{5}$$

Here the coefficients $\lambda_n, f_{n+1,n}, \ldots, f_{n+1,0}$ are so chosen that $(Zr^{(i)}, r^{(j)}) = 0$ for $i < j$. In the symmetrizable case if Z and ZA are SPD ORTHORES reduces to ORTHORES(1). ORTHORES(1) was given by Engeli, et al [3]. An alternative form of ORTHORES(1) was given by Concus, Golub, and O'Leary [1].

As shown by Jea [7], each of the three forms of the IGCG method has the property that for some $t \leq N$ if the method does not break down within t iterations then we have convergence in the sense that $u^{(t)} = \bar{u} = A^{-1} b$. Furthermore ORTHOMIN converges if and only if ORTHORES

converges and if both converge then ORTHODIR converges. Sufficient conditions for convergence are that ZA be positive real (PR), (i.e., ZA + $(ZA)^T$ is SPD), for ORTHODIR, and that ZA and Z be PR for ORTHOMIN and ORTHORES. Elman [2] has shown that if A is PR then ORTHOMIN(s) converges.

Unfortunately, for the IGCG method the determination of $u^{(n+1)}$ requires, in general, some information from each of the previous n iterations. This can be prohibitive in terms of computer time and storage. One remedy for this is to use truncated schemes. Unfortunately most of the theoretical properties of the idealized schemes do not carry over to the truncated schemes. Nevertheless, numerical experiments reported by Elman [2], by Young and Jea [12], and by Jea [6] indicate that ORTHOMIN(s) and ORTHORES(s) with small s ($2 \leq s \leq 4$) are effective in many cases. Much more numerical and experimental work remains to be done.

The emphasis of the talk is to study cases where, as in the symmetrizable case, the three forms of the IGCG method are equivalent to the corresponding

truncated versions. We define Condition I as follows. The nonsingular matrix H satisfies Condition I with respect to A if $HA = A^TH$. It is shown that such a matrix H always exists. However, in most cases it is not feasible to find H. However, if A is symmetric (through not necessarily SPD) one can let H = I. Similarly if A is similar to a symmetric matrix and if a matrix C is available such that CAC^{-1} is symmetric then H can be found. In any case, given an H such that Condition I holds, with Z = H or $Z = A^TH$ ORTHODIR, ORTHOMIN and ORTHORES reduce to ORTHODIR(2), ORTHOMIN(1), and ORTHORES(1) respectively. If H is symmetric it is better to let $Z = A^TH$ for ORTHODIR and ORTHOMIN while Z = H should be used for ORTHORES.

While it is seldom possible to find a matrix H satisfying Condition I, nevertheless, there is one important case where H can be found. We consider the expanded system

(6) $$\textcircled{A}\ \textcircled{u} = \textcircled{b}$$

where

$$\text{(7)}\qquad \mathbb{A} = \begin{bmatrix} A & 0 \\ 0 & A^T \end{bmatrix}, \quad \textcircled{u} = \begin{bmatrix} u \\ \tilde{u} \end{bmatrix}, \quad \textcircled{b} = \begin{bmatrix} b \\ \tilde{b} \end{bmatrix}$$

Evidently $\textcircled{u}$ satisfies (6) if and only if u satisfies (1) and $\tilde{u}$ satisfies the fictitious system

$$\text{(8)}\qquad A^T\tilde{u} = \tilde{b}$$

Moreover one can show that $\textcircled{H}\textcircled{\mathbb{A}} = \textcircled{\mathbb{A}}^T\textcircled{H}$ where

$$\text{(9)}\qquad \textcircled{H} = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$$

If we apply the IGCG method to (6) with $\textcircled{Z} = \textcircled{H}$ we get three alternative forms of the Lanczos method, (Lanczos [9,10]) corresponding to the three forms of the IGCG method. We refer to these three forms as the Lanczos/ ORTHODIR, Lanczos/ORTHOMIN, and the Lanczos/ORTHORES methods. The second form is the same as the biconjugate gradient method of Fletcher [4,5]. Convergence results are given for the three forms of the Lanczos method, which are analogus to those obtained for the IGCG method. Justification is given for choosing the Lanczos/ORTHODIR form in preference to the more frequently used Lanczos/ORTHOMIN and Lanczos/ORTHORES forms.

REFERENCES

1. P. Concus, G. H. Golub and D. P. O'Leary, A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations, in Sparse Matrix Computation (J. R. Bunch and and D. J. Rose, eds.) Academic Press, New York, 309-332, (1976).

2. Howard C. Elman, Iterative methods for large, sparse, nonsymmetric systems of linear equations, Research Report 229, Yale University, Department of Computer Sciences, April 1982.

3. M. Engeli, T. Ginsburg, H. Rutishauser and E. Stiefel, Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems, Mitt. Inst. Angew. Math, ETH, Zurich, Nr. 8, Basel-Stuttgart, (1959).

4. R. Fletcher, Conjugate gradient methods for indefinite systems, Proc. of the Dundee Biennial Conference on Numerical Analysis (G. A. Watson, ed.) pp. 73-89, Springer-Verlag, Berlin and New York (1975).

5. R. Fletcher, Conjugate gradient methods for indefinite systems, Lecture Notes in Mathematics 506, Springer-Verlag, Berlin and New York (1976).

6. M. R. Hestenes, The conjugate-gradient method for solving linear systems in Numerical Analysis, Vol. VI (J. Curtiss, ed.), McGraw-Hill, New York, (1956).

7. K. C. Jea, Generalized conjugate gradient acceleration of iterative methods, Ph. D. Thesis, Department of Mathematics, The University of Texas at Austin (1982); also CNA-176, Center for Numerical Analysis, The University of Texas at Austin (1982).

8. K. C. Jea and David M. Young, On the simplification of generalized conjugate gradient methods for nonsymmetrizable linear systems, to appear in The Journal of Linear Algebra and its Applications.

9. C. Lanczos, An iterative method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. Nat. Bur. Standards 45: 255-282, (1950).

10. C. Lanczos, Solution of systems of linear equations by minimized iterations, J. Res. Nat. Bur. Standards 49: 33-53, (1952).

11. D. M. Young and K. C. Jea, Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods, J. of Linear Algebra and Its Applications 34: 159-194, (1980).

12. D. M. Young and K. C. Jea, Generalized conjugate gradient acceleration of iterative methods, Part II; the Nonsymmetrizable Case, CNA-163, Center for Numerical Analysis, The University of Texas at Austin (1981).