ON THE LINEAR LEAST SQUARES PROBLEM WITH A QUADRATIC CONSTRAINT

by

Walter Gander

STAN-CS-78-697
November 19 78

C O M P U T E R S C I E N C E D E P A R T M E N T
School of Humanities and Sciences
STANFORD UNIVERSITY

On the Linear Least Squares Problem

with a Quadratic Constraint

by

Walter Gander*

* Neu-Technikum Buchs, CH-9470 Buchs, Switzerland.

Research supported in part by National Science Foundation Grant No. MCS78-17697.

On the Linear Least Squares Problem

with a Quadratic Constraint

by

Walter Gander */

Abstract.

In this paper we present the theory and practical computational aspects of the linear least squares problem with a quadratic constraint. New theorems characterizing properties of the solutions are given and extended for the problem of minimizing a general quadratic function subject to a quadratic constraint. For two important regularization methods we formulate dual equations which proved to be very useful for the application of smoothing of datas. The resulting algorithm is a numerically stable version of an algorithm proposed by Rutishauser. We show also how to choose a third order iteration method to solve the secular equation. However we are still far away from a foolproof machine independent algorithm.

*/ Computer Science Department, Stanford University, Stanford, Calif. 94305.
On leave from Neu-Technikum Ruchs, Switzerland.
The author has been supported by the Swiss National Science Foundation, Grant No. 5'521'330'61517.

To Maurice

Acknowledgement.

This work was done during my Stanford year 1977/78. I am greatly indebted to Prof. G. Golub for his valuable comments end suggestions, for his hospitality and for creating the extraordinary spirit of Serra House that attracts in one year most numerical analysts of the world. My stay would not have been possible without the support of Prof. P. Henrici, one of my "godfathers" for the Swiss NSF. I would like to thank him for his continuous interest in my work and for his encouragement. I owe him much of my mathematical education. Also in Switzerland I have to thank Prof. J. Marti, my second "godfather", who happened to work at a similar problem. I am indebted to Prof. Ch. van Loan, Cornell University, for his careful reading of the manuscript end for his various suggestions helping to improve my english. Thanks also to the "Forschungskommission" of ETHZ and the Swiss National Science Foundation who gave me the grant that enabled my stay at Stanford. Finally I have to thank my wife Heidi and my parents-in-law, A. and E. Wolf, for their non-mathematical assistance.

Stanford, Fall 1978
Walter Gander

Table of Contents

0. Introduction.

In this paper we consider least squares problems with a quadratic constraint. The matrices and vectors will be real and we use capital letters $\underline{A}$, $\underline{B}$, . . . for matrices and small letters $\underline{a}, \underline{b}$, . . . for vectors. $\|\ \|_{II}$ will be used for the euclidian vectornorm.

Given $\underline{A}$, $\underline{C}$, $\underline{b}$, $\underline{d}$ and a number $\alpha \geq 0$ we consider the problem to find x such that

$$\left.\begin{aligned} &\|\underline{Ax} - \underline{b}\| = \min \\ \text{subject to}\quad &\|\underline{Cx} - \underline{d}\| \leq \alpha \quad . \end{aligned}\right\} \tag{0.1}$$

This problem is a generalization of the least square problem with equality constraints

$$\left.\begin{aligned} &\|\underline{Ax} - \underline{b}\| = \min \\ \text{subject to}\quad &\underline{Cx} = \underline{d} \end{aligned}\right\} \tag{0.2}$$

since for $\alpha = 0$ every solution of (0.1) is **also** a solution of (0.2).

The motivation why to consider (0.1) rather than (0.2) is explained best by the following example. Let's assume we are given m values of a function f

$$y_i = f(t_i) , \quad i = 1,\ldots,m$$

and we seek the coefficients of a **polynomial**

$$p(t) = \sum_{i=0}^{n-1} x_i t^i$$

that approximates f .

If we insist that p interpolate the given data, then we have to solve the system (m equations and n unknowns)

$$\underline{\underline{A}}\underline{x} = \underline{y} \tag{0.3}$$

with $a_{ij} = t_i^j$. If $m > n$, (0.3) may have no solution. If $m < n$, the solution is not unique. If $m = n$, there is a unique solution if $t_i \neq t_j$, $i \neq j$. However it is well known that $\underline{\underline{A}}$ is ill conditioned which means that x is difficult to compute accurately and that usually the norm of x will be large. The polynomial p with large coefficient will be useless since cancellation will affect its evaluation. It may therefore be better not to interpolate but to approximate the datas in the least squares sense. This leads to the unconstrained least squares problem

$$\|\underline{\underline{A}}\underline{x} - \underline{y}\| = \min . \tag{0.4}$$

This problem has for $m \geq n$ and if A has full rank a unique solution. If $\underline{\underline{A}}$ is rank deficient (0.4) has infinitely many solutions and therefore one usually looks for the solution with minimum norm

$$\left.\begin{aligned} &\min \|\underline{x}\| \\ \text{subject to}\quad &\|\underline{\underline{A}}\underline{x} - \underline{y}\| = \min . \end{aligned}\right\} \tag{0.5}$$

The solution of (0.5) is given explicitly by

$$\underline{x} = \underline{\underline{A}}^{+}\underline{y}$$

where $\underline{\underline{A}}^{+}$ is the pseudoinverse of A . However the solution of (0.4) [35] and (0.5) may also suffer having a large norm. The problem is then said to be ill posed and we consider two ways to regularize [42] the solution. We can prescribe a bound for $\|\underline{x}\|$ and thus look for the solution of

$$\left.\begin{aligned} &\|\underline{\underline{A}}\underline{x} - \underline{y}\| = \min \\ \text{subject to}\quad &\|\underline{x}\| \leq \alpha . \end{aligned}\right\} \tag{0.6}$$

Alternatively we may prefer that the deviation from the given data points be bounded. This leads to

$$\left.\begin{aligned} &\|\underline{x}\| = \min \\ \text{subject to}\quad &\|\underline{\underline{A}}\underline{x} - \underline{y}\| \leq \alpha . \end{aligned}\right\} \tag{0.7}$$

The question how to choose α is not simple to answer. It may be necessary to use some statistical tools to estimate α [17]. we observe that (0.6) and (0.7) are least squares problems with a quadratic constraint. The degree of the polynomial (n-1) may be chosen independently of the number of given data points. We may wish to compute a low degree polynomial ($m \gg n$) or a polynomial with large degree $m < n$ but with small coefficients. Problems (0.6) and (0.7) have a unique solution (if it exists) and we will show how to compute it efficiently.

Returning to our example, let's consider a partitioning of the data points in two sets. We may ask for a polynomial that interpolates the datas of the first set exactly. This leads to a least squares problem with equality constraints:

$$\left.\begin{aligned} &\|\underline{\underline{C}}\underline{x} - \underline{y}_2\| = \min \\ \text{subject to}\quad &\underline{\underline{B}}\underline{x} = \underline{y}_1 \end{aligned}\right\} \tag{0.8}$$

where $\underline{y}_1$ contains the function values of the first set. Instead of interpolating on the first set we could ask for a bound of the deviation and get again problem (0.1):

$$\left.\begin{aligned} &\|\underline{\underline{C}}\underline{x} - \underline{y}_2\| = \min \\ \text{subject to}\quad &\|\underline{\underline{B}}\underline{x} - \underline{y}_1\| \leq \alpha . \end{aligned}\right\} \tag{0.9}$$

In contrast to (0.6) and (0.7), (0.9) may not have a unique solution. As we shall see the solution is unique if and only if the nullspaces of $\underline{C}$ and $\underline{B}$ intersect trivially.

1. Basic Definitions and Remarks.

The following notation will be used:

Problem (P1):

$$\|\underline{A}\underline{x} - \underline{b}\| = \min \tag{1.1}$$

subject to

$$\|\underline{C}\underline{x} - \underline{d}\| \leq \alpha \quad . \tag{1.2}$$

If we have $\|\underline{C}\underline{x} - \underline{d}\| = \alpha$ instead of (1.2) we will refer to the problem as (P1E). Two special cases of (P1) will be considered:

Problem (P2): like (P1) but $C = \underline{I}$, $d = \underline{0}$.

Problem (P3): like (P1) but $A = \underline{I}$, $b = \underline{0}$.

Finally (P2E) and (P3E) will be the corresponding problems with equality sign in the constraint.

The solution of (P1) is a stationary point of the Lagrange function (with the Lagrange multiplier λ)

$$L(\underline{x},\lambda) = \|\underline{A}\underline{x} - \underline{b}\|^2 + \lambda\{\|\underline{C}\underline{x} - \underline{d}\|^2 - \alpha^2\}$$

and therefore a solution of $\frac{\partial L}{\partial \underline{x}} = \underline{0}$ and $\frac{\partial L}{\partial \lambda} = 0$, which are the "normal equations":

$$(\underline{A}^T\underline{A} + \lambda\, \underline{C}^T\underline{C})\underline{x} = \underline{A}^T\underline{b} + \lambda\, \underline{C}^T\underline{d} \tag{1.3}$$

$$\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2 \quad . \tag{1.4}$$

If the matrix $\underline{A}^T\underline{A} + \lambda\, \underline{C}^T\underline{C}$ is nonsingular, then we can define

$$f(\lambda) := \|\underline{C}\ \underline{x}(\lambda) - \underline{d}\|^2 \tag{1.5}$$

where $\underline{x}(\lambda)$ is the solution of (1.3). We will call f the "length

function". To determine a solution we have to solve the "secular equation"

$$f(h) = \alpha^2 \quad . \tag{1.6}$$

Finally we observe that for $\lambda > 0$ equations (1.3) and (1.4) are the normal equations of the least squares problem

$$\left\| \begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\ \underline{C} \end{pmatrix} \underline{x} - \begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\ \underline{d} \end{pmatrix} \right\| = \min \quad .$$

A useful tool for the analysis of problems (P2) and (P3) is the singular value decomposition (SVD) [14] and its generalization (BSVD) [45] for (P1). These decompositions can also be used for the practical computations. However for the problems (P1E), (P2E) and (P3E) there are less expensive ways [8]. In some applications [33], $\underline{A}$ and C are band matrices and (P1) may be solved efficiently without transformations as we shall show.

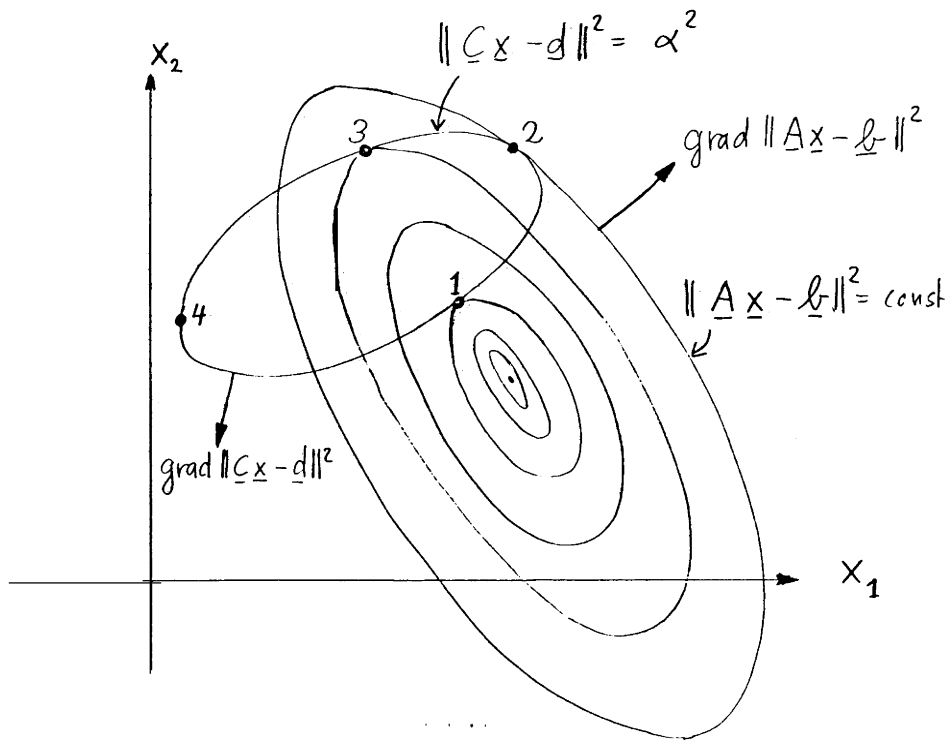
2. The Least Squares Problem with a Quadratic Constraint.

Let A be an $(m \times n)$ -matrix, C a $(p \times n)$ -matrix, b an m -vector, d a p -vector, and α a positive number.

We consider the problem to find an n-vector x so that

$$\left.\begin{aligned} &\|\underline{A}\underline{x} - \underline{b}\| = \min \\ \text{subject to} \quad &\|\underline{C}\underline{x} - \underline{d}\| = \alpha \end{aligned}\right\} \tag{P1E}$$

For $n = 2$ we can interpret this problem geometrically. The level lines of $\|\underline{A}\underline{x} - \underline{b}\|^2 = \text{const}$ are ellipses centered at $\underline{A}^+\underline{b}$. The constraints $\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2$ is also an ellipse:



We are looking for a point $\underline{x}$ on the ellipse $\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2$ which has the smallest value of $\|\underline{A}\underline{x} - \underline{b}\|^2$. Clearly it is the point 1 . At that point the gradients of the two functions

$$\|\underline{C}\underline{x} - \underline{d}\|^2 \quad , \quad \|\underline{A}\underline{x} - \underline{b}\|^2$$

are parallel, i.e., some λ exist so that

$$\text{grad } \|\underline{A}\underline{x} - \underline{b}\|^2 = -\lambda \text{ grad } \|\underline{C}\underline{x} - \underline{d}\|^2 \quad .$$

If we rearrange this equation we have

$$(\underline{A}^T\underline{A} + \lambda \underline{C}^T\underline{C})\underline{x} = \underline{A}^T\underline{b} + \lambda \underline{C}^T\underline{d}$$

which is equation (1.3) that we obtained using the technique of the Lagrange multipliers.

We observe that there are three other points for which the gradients of the two functions are parallel (2,3,4) . However for these three points λ is negative since the gradients have the same directions.

We can think of problem (PlE) as blowing up a balloon centered at $\underline{A}^+\underline{b}$ which has an ellipsoid shape until it touches the ellipsoid $\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2$.

If $\underline{A}^+\underline{b}$ is outside of $\|\underline{C}\underline{x} - \underline{d}\|^2$ then at the point of touch the gradients will have opposite sign, i.e., $\lambda > 0$. However if $\underline{A}^+\underline{b}$ is inside the gradients have the same sign and $\lambda < 0$.

For the problems with inequality cons-taints (P1), (P2) and (P3) we have only to consider the case that $\underline{A}^+\underline{b}$ is outside of $\|\underline{C}\underline{x} - \underline{d}\|^2$. If it is inside then $\underline{x} = \underline{A}^+\underline{b}$ solves the problem.

2.1 Characterization of the Solution.

The solution of (PlE) is among the solutions $(\underline{x},\lambda)$ of the normal equations (see Section 1):

$$\left.\begin{aligned} (\underline{A}^T\underline{A} + \lambda \underline{C}^T\underline{C})\underline{x} &= \underline{A}^T\underline{b} + \lambda \underline{C}^T\underline{d} \\ \|\underline{C}\underline{x} - \underline{d}\|^2 &= \alpha^2 \quad . \end{aligned}\right\} \tag{2.1}$$

The following theorem compares two solutions of these equations:

Theorem 2.1. If $(\underline{x}_1,\lambda_1)$ and $(\underline{x}_2,\lambda_2)$ are solutions of the normal equations (2.1.), then

$$\|\underline{A}\underline{x}_2 - \underline{b}\|^2 - \|\underline{A}\underline{x}_1 - \underline{b}\|^2 = \frac{\lambda_1 - \lambda_2}{2} \|\underline{C}(\underline{x}_1 - \underline{x}_2)\|^2 \quad . \tag{2.2}$$

Proof. Since $(\underline{x}_1,\lambda_1)$, $(\underline{x}_2,\lambda_2)$ are solutions of (2.1) we have

$$\underline{A}^T\underline{A}\underline{x}_1 - \underline{A}^T\underline{b} = -\lambda_1 \underline{C}^T\underline{C}\underline{x}_1 + \lambda_1 \underline{C}^T\underline{d} \tag{2.3}$$

$$\underline{A}^T\underline{A}\underline{x}_2 - \underline{A}^T\underline{b} = -\lambda_2 \underline{C}^T\underline{C}\underline{x}_2 + \lambda_2 \underline{C}^T\underline{d} \tag{2.4}$$

$\underline{x}_2^T$ (2.4) - $\underline{x}_1^T$ (2.3) gives

$$\|\underline{A}\underline{x}_2\|^2 - \|\underline{A}\underline{x}_1\|^2 - \underline{b}^T\underline{A}(\underline{x}_2 - \underline{x}_1)$$

$$= \lambda_1(\|\underline{C}\underline{x}_1\|^2 - \underline{d}^T\underline{C}\underline{x}_1) - \lambda_2(\|\underline{C}\underline{x}_2\|^2 - \underline{d}^T\underline{C}\underline{x}_2) \tag{2.5}$$

$\underline{x}_2^T$ (2.3) - $\underline{x}_1^T$ (2.4) gives

$$-\underline{b}^T(\underline{A}\underline{x}_2 - \underline{x}_1) = \lambda_1(-\underline{x}_2^T \underline{C}^T\underline{C}\underline{x}_1 + \underline{d}^T\underline{C}\underline{x}_2) - \lambda_2(-\underline{x}_1^T \underline{C}^T\underline{C}\underline{x}_1 + \underline{d}^T\underline{C}\underline{x}_1) \ . \tag{2.6}$$

Observe that

$$\|A\underline{x}_2 - \underline{b}\|^2 - \|A\underline{x}_1 - \underline{b}\|^2 = \|A\underline{x}_2\|^2 - \|A\underline{x}_1\|^2 - 2\underline{b}^T A(\underline{x}_2 - \underline{x}_1) \quad .$$

So that if we add (2.5) - (2.6) we get

$$\|A\underline{x}_2 - \underline{b}\|^2 - \|A\underline{x}_1 - \underline{b}\|^2$$

$$= \lambda_1\{\|C\underline{x}_1\|^2 - \underline{d}^T C\underline{x}_1 - \underline{x}_1^T C^T C\underline{x}_2 + \underline{d}^T C\underline{x}_2\}$$

$$- \lambda_2\{\|C\underline{x}_2\|^2 - \underline{d}^T C\underline{x}_2 - \underline{x}_1^T C^T C\underline{x}_2 + \underline{d}^T C\underline{x}_1\} \quad . \qquad (2.7)$$

Now we have

$$\|C\underline{x}_1 - \underline{d}\|^2 = \|C\underline{x}_2 - \underline{d}\|^2 = \alpha^2$$

$$\Rightarrow \|C\underline{x}_1\|^2 - 2\underline{d}^T C\underline{x}_1 + \|\underline{d}\|^2 = \|C\underline{x}_2\|^2 - 2\underline{d}^T C\underline{x}_2 + \|\underline{d}\|^2$$

$$\Rightarrow \|C\underline{x}_1\|^2 - \underline{d}^T C\underline{x}_1 + \underline{d}^T C\underline{x}_2 = \|C\underline{x}_2\|^2 - \underline{d}^T C\underline{x}_2 + \underline{d}^T C\underline{x}_1 \quad . \qquad (2.8)$$

From (2.8) we conclude that the factors of λ_1 and λ_2 in (2.7) are the same. Therefore they also equal their arithmetic mean which is

$$\frac{1}{2}\{\|C\underline{x}_1\|^2 - 2\underline{x}_1^T C^T C\underline{x}_2 + \|C\underline{x}_2\|^2\}$$

$$= \frac{1}{2}\|C(\underline{x}_1 - \underline{x}_2)\|^2 \quad . \quad \square$$

Corollary 2.1. The solution of (PLE) is the solution $\underline{x}(\lambda)$ of the normal equations (2.1) with the largest λ .

Proof. From (2.2) we have that if $\lambda_1 > \lambda_2$, then $\|A\underline{x}_2 - \underline{b}\|^2 > \|A\underline{x}_1 - \underline{b}\|^2$. $\square$

The next theorem gives a result very similar to Theorem 2.1.

Theorem 2.2. Assume $(\underline{x}_1,\lambda_1)$ and $(\underline{x}_2,\lambda_2)$ are solutions of the normal equations (2.1). Assume that $|\lambda_1| + |\lambda_2| \neq 0$. Then

$$\|A\underline{x}_2 - \underline{b}\|^2 - \|A\underline{x}_1 - \underline{b}\|^2 = \frac{A_2 - A_1}{\lambda_2 + \lambda_1}\|A(\underline{x}_1 - \underline{x}_2)\|^2 \, .$$

Proof. From $(A^TA + \lambda C^TC)\underline{x} = A^T\underline{b} + \lambda C^T\underline{d}$ we have

$$\lambda_1 C^T C\underline{x}_1 - \lambda_1 C^T\underline{d} = -A^TA\underline{x}_1 + A^T\underline{b} \qquad (2.8)$$

$$\lambda_2 C^T C\underline{x}_2 - \lambda_2 C^T\underline{d} = -A^TA\underline{x}_2 + A^T\underline{b} \quad . \qquad (2.9)$$

$\lambda_1\underline{x}_1^T$ (2.9) - $\lambda_2\underline{x}_2^T$ (2.8) gives

$$\lambda_1\lambda_2\{(\underline{x}_2 - \underline{x}_1)^T C^T\underline{d}\} = (\lambda_2 - \lambda_1)\underline{x}_1^T A^TA\underline{x}_2 + (\lambda_1\underline{x}_1 - \lambda_2\underline{x}_2)^T A^T\underline{b} \quad . \qquad (2.10)$$

$\lambda_1\underline{x}_2^T$ (2.9) - $\lambda_2\underline{x}_1^T$ (2.8) gives

$$\lambda_1\lambda_2\{\|C\underline{x}_2\|^2 - \|C\underline{x}_1\|^2 + (\underline{x}_1 - \underline{x}_2)C^T\underline{d}\}$$

$$= \lambda_2\|A\underline{x}_1\|^2 - \lambda_1\|A\underline{x}_2\|^2 + (\lambda_1\underline{x}_2 - \lambda_2\underline{x}_1)^T A^T\underline{b} \quad . \qquad (2.11)$$

Observe that

$$0 = \|C\underline{x}_2 - \underline{d}\|^2 - \|C\underline{x}_1 - \underline{d}\|^2$$

$$= \|C\underline{x}_2\|^2 - \|C\underline{x}_1\|^2 + 2(\underline{x}_1 - \underline{x}_2)^T C^T\underline{d} \quad .$$

So that if we subtract (2.11) - (2.10) we get

. . .

$$0 = \lambda_2\|\underline{A}\underline{x}_1\|^2 - \lambda_1\|\underline{A}\underline{x}_2\|^2 + (\lambda_1\underline{x}_2 - \lambda_2\underline{x}_1 - \lambda_1\underline{x}_1 + \lambda_2\underline{x}_2)^T \underline{A}^T\underline{b}$$

$$+ (\lambda_1 - \lambda_2)\, \underline{x}_1^T \underline{A}^T\underline{A}\underline{x}_2 \quad ,$$

or by rearranging

$$\lambda_1\{\|\underline{A}\underline{x}_2\|^2 - \underline{x}_2^T \underline{A}^T\underline{b} + \underline{x}_1^T \underline{A}^T\underline{b} - \underline{x}_1^T \underline{A}^T\underline{A}\underline{x}_2\}$$

$$= \lambda_2\{\|\underline{A}\underline{x}_1\|^2 - \underline{x}_1^T \underline{A}^T\underline{b} + \underline{x}_2^T \underline{A}^T\underline{b} - \underline{x}_1^T \underline{A}^T\underline{A}\underline{x}_2\} \quad . \qquad (2.12)$$

Now the { } on the left hand side of (2.12) is

$$\tfrac{1}{2}\|\underline{A}\underline{x}_2\|^2 + \tfrac{1}{2}\{\|\underline{A}\underline{x}_2\|^2 - 2\underline{x}_1^T \underline{A}^T\underline{A}\underline{x}_2 + \|\underline{A}\underline{x}_1\|^2\}$$

$$- \tfrac{1}{2}\|\underline{A}\underline{x}_1\|^2 + \underline{x}_1^T \underline{A}^T\underline{b} - \underline{x}_2^T \underline{A}^T\underline{b} + \tfrac{1}{2}\|\underline{b}\|^2 - \tfrac{1}{2}\|\underline{b}\|^2$$

$$= \tfrac{1}{2}\{\|\underline{A}\underline{x}_2 - \underline{b}\|^2 - \|\underline{A}\underline{x}_1 - \underline{b}\|^2 + \|\underline{A}(\underline{x}_2 - \underline{x}_1)\|^2\} \quad .$$

The right hand side of (2.12) simplifies analogously and by rearranging we obtain:

$$(\lambda_1 + \lambda_2)\{\|\underline{A}\underline{x}_2 - \underline{b}\|^2 - \|\underline{A}\underline{x}_1 - \underline{b}\|^2\} = (\lambda_2 - \lambda_1)\|\underline{A}(\underline{x}_2 - \underline{x}_1)\|^2 \quad . \qquad (2.13)$$

We now prove by contradiction that $\lambda_1 + \lambda_2 \neq 0$. Assume $(\underline{x}_1, \lambda)$ and $(\underline{x}_2, -\lambda)$ were solutions of the normal equations (2.1) with $A > 0$. Then by (2.13) we would have

$$\|\underline{A}(\underline{x}_2 - \underline{x}_1)\|^2 = 0$$

$$\Rightarrow \underline{A}\underline{x}_2 = \underline{A}\underline{x}_1$$

$$\Rightarrow \|\underline{A}\underline{x}_2 - \underline{b}\|^2 = \|\underline{A}\underline{x}_1 - \underline{b}\|^2 \quad .$$

But by Corollary 2.1,

$$\lambda_1 = A > -A = \lambda_2 \Rightarrow \|\underline{A}\underline{x}_2 - \underline{b}\|^2 > \|\underline{A}\underline{x}_1 - \underline{b}\|^2 \quad .$$

Therefore $\lambda_1 + \lambda_2 \neq 0$ and we may divide in (2.13). □

If we combine both results from Theorem 2.1 and Theorem 2.2 we have

Corollary 2.2. Let $(\underline{x}_1, \lambda_1)$ and $(\underline{x}_2, \lambda_2)$ be two solutions of the normal equations (2.1). If $|\lambda_1| + |\lambda_2| \neq 0$, then

$$-\frac{\lambda_1 + \lambda_2}{2}\|\underline{C}(\underline{x}_1 - \underline{x}_2)\|^2 = \|\underline{A}(\underline{x}_1 - \underline{x}_2)\|^2 \quad . \qquad (2.14)$$

From (2.14) we see immediately:

Corollary 2.3. The normal equations (2.1) have at most one solution $(\underline{x}^*, \lambda^*)$ with $A^* > 0$. For every other solution $(\underline{x}, \lambda)$ we have

$$A < -\lambda^* \quad .$$

The next theorem gives conditions for a unique solution of (PlE).

Theorem 2.3. The solution x of (PlE) is unique (if it exists) if

$$NS(\underline{A}) \cap NS(\underline{C}) = \{\underline{0}\}$$

and

$$\lambda \neq -\mu_i$$

where $(\underline{x}, \lambda)$ is a solution of the normal equation (2.1) and μ_i is an eigenvalue of the generalized eigenvalue problem

$$\det(\underline{A}^T\underline{A} - \mu\, \underline{C}^T\underline{C}) = 0 \quad .$$

Proof. The proof is by contradiction. Assume $(\underline{x}_1,\lambda_1)$ and $(\underline{x}_2,\lambda_2)$ are solutions of the normal equations (2.1) which also solve (PlE).

If $\lambda_1 \neq \lambda_2$, then we must have

$$\|\underline{A}\underline{x}_1 - \underline{b}\|^2 = \|\underline{A}\underline{x}_2 - \underline{b}\|^2 = \min \|\underline{A}\underline{x} - \underline{b}\|^2 .$$

Theorem 2.1 implies

$$\|\underline{C}(\underline{x}_1 - \underline{x}_2)\|^2 = 0 \Rightarrow \underline{C}(\underline{x}_1 - \underline{x}_2) = \underline{0} . \tag{2.15}$$

While Theorem 2.2 gives

$$\|\underline{A}(\underline{x}_1 - \underline{x}_2)\|^2 = 0 \Rightarrow \underline{A}(\underline{x}_1 - \underline{x}_2) = \underline{0} . \tag{2.16}$$

Now if $\underline{x}_1 \neq \underline{x}_2$ then (2.15) and (2.16) show that $\underline{A}$ and $\underline{C}$ have non-trivially intersecting nullspaces, which is a contradiction.

Therefore we must have $\lambda_1 = \lambda_2 = A$. But in this case we then have

$$(\underline{A}^T\underline{A} + \lambda\ \underline{C}^T\underline{C})(\underline{x}_1 - \underline{x}_2) = 0 .$$

If $\underline{x}_1 \neq \underline{x}_2$ then $\lambda = -\mu_i$ which is also a contradiction. Therefore we must have $\lambda_1 = \lambda_2$ and $\underline{x}_1 = \underline{x}_2$. □

$\underline{A}$ and $\underline{C}$ have a trivial intersection of their nullspaces if and only if

$$\operatorname{rank}\begin{pmatrix}\underline{A}\\ \underline{C}\end{pmatrix} = n .$$

Therefore a necessary condition for a unique solution is

$$m+p > n$$

which means that we must have "enough" equations to determine $\underline{x}$.

We shall assume in the following (till the end of the paper) that $NS(\underline{A}) \cap NS(\underline{C}) = \{\underline{0}\}$. In Section 2.2 we shall analyze the existence of

solutions of the normal equations. We shall see that (PlE) has a solution if the constraint

$$\|\underline{C}\underline{x} - \underline{d}\| = \alpha^2$$

is feasible, i.e., if

$$\min_{x} \|\underline{C}\underline{x} - \underline{d}\|^2 = \|(\underline{C}\ \underline{C}^+ - \underline{I})\underline{d}\|^2 < \alpha^2 .$$

2.2 The Solutions of the Normal Equations.

In this section we shall discuss the solutions of

$$(\underline{A}^T\underline{A} + \lambda\ \underline{C}^T\underline{C})\underline{x} = \underline{A}^T\underline{b} + A\ \underline{C}^T\underline{d} \tag{2.17a}$$

$$\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2 . \tag{2.17b}$$

We note that we have assumed that $NS(\underline{A}) \cap NS(\underline{C}) = \{\underline{0}\}$.

If $\lambda \neq -\mu_i$ where $\mu_i \geq 0$ is an eigenvalue of the eigenvalue problem $\det(\underline{A}^T\underline{A} - \mu\ \underline{C}^T\underline{C}) = 0$ then

$$\underline{x}(\lambda) = (\underline{A}^T\underline{A} + \lambda\ \underline{C}^T\underline{C})^{-1}(\underline{A}^T\underline{b} + \lambda\ \underline{C}^T\underline{d}) . \tag{2.18}$$

If we now choose λ so that the secular equation is satisfied:

$$f(A) := \|\underline{C}\ \underline{x}(\lambda) - \underline{d}\|^2 = \alpha^2 \tag{2.19}$$

then $(\underline{x}(\lambda),\lambda)$ is a solution of (2.17). We first discuss some properties of the length function f :

Lemma 2.1. If f is defined by (2.19), (2.18), then

(1) f is a rational function defined on $\mathbb{R} - \{-\mu_i \mid \det(\underline{A}^T\underline{A} - \mu_i\ \underline{C}^T\underline{C}) = 0\}$,

(2) $f(\lambda) \equiv \|\underline{d}\|^2 = \text{const}$ if $\underline{A}^T\underline{b} = \underline{0}$ and $\underline{C}^T\underline{d} = \underline{0}$,

(3) f has at least one and at *most* n poles *for sane* $\lambda = -\mu_i$,

(4) $f(\lambda) > 0$, $\lim_{A \to +\infty} f(\lambda) = \|(\underline{C}\,\underline{C}^{+} - \underline{I})\underline{d}\|^2$.

(5) $f'(A) < 0$ for $0 < \lambda < \infty$.

proof. We use the generalized singular value decomposition BSVD [45] which gives the decomposition

$$\left.\begin{aligned} \underline{U}^T\underline{A}\underline{X} &= \underline{D}_A = \mathrm{diag}(\alpha_1,\ldots,\alpha_n)\ ,\ \alpha_i \geq 0 \\ \underline{V}^T\underline{C}\underline{X} &= \underline{D}_C = \mathrm{diag}(\gamma_1,\ldots,\gamma_q)\ ,\ \gamma_i \geq 0\ ,\ q = \min(n,p) \end{aligned}\right\} \qquad (2.20)$$

where $\underline{U}$ $(m \times m)$ and $\underline{V}$ $(p \times p)$ are orthogonal, $\underline{X}$ $(n \times n)$ is nonsingular, and $\underline{D}_A$, $\underline{D}_C$ are diagonal matrices with $\gamma_1 \geq \ldots \geq \gamma_\lambda$. The decomposition exists only if $m > n$ which we can assume since we can add in

$$\|\underline{A}\underline{x} - \underline{b}\|^2 = \left\| \begin{pmatrix} \underline{A} \\ \underline{0} \end{pmatrix}\underline{x} - \begin{pmatrix} \underline{b} \\ \underline{0} \end{pmatrix} \right\|^2$$

zero rows in $\underline{A}$ and zero elements in $\underline{b}$ without changing the solutions.

If $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r > \gamma_{r+1} = \cdots = \gamma_q = 0$, then the eigenvalue of $\det(\underline{A}^T\underline{A} - \mu\,\underline{C}^T\underline{C}) = 0$ are given by

$$\mu_i = \frac{\alpha_i^2}{\gamma_i^2} \quad , \quad i = 1,\ldots,r\ .$$

Because we are assuming that $NS(\underline{A}) \cap NS(\underline{C}) = \{\underline{0}\}$ we have

$$\alpha_i > 0 \quad \text{for} \quad i = r+1,\ldots,n\ . \quad [45]$$

By substituting (2.20) into (2.16) we have

$$\underline{X}^{-T}(\underline{D}_A^T\,\underline{D}_A + \lambda\,\underline{D}_C^T\,\underline{D}_C)\underline{X}^{-1}\,\underline{x} = X^{-T}\,\underline{D}_A^T\,\underline{U}^T\underline{b} + A\,X^{-T}\,\underline{D}_C^T\,\underline{V}^T\underline{d}$$

and putting

$$\underline{y} := \underline{X}^{-1}\underline{x} \quad , \quad \underline{c} := \underline{U}^T\underline{b} \quad , \quad \underline{e} := \underline{V}^T\underline{d}$$

we get

$$(\underline{D}_A^T\,\underline{D}_A + \lambda\,\underline{D}_C^T\,\underline{D}_C)\underline{y} = \underline{D}_A^T\,\underline{c} + \lambda\,\underline{D}_C^T\,\underline{e}\ . \qquad (2.21)$$

Now

$$f(\lambda) = \|\underline{C}\,\underline{x}(\lambda) - \underline{d}\|^2 = \|\underline{D}_C\,\underline{y}(\lambda) - \underline{e}\|^2\ ,$$

which is in components:

$$f(A) = \sum_{i=1}^{r}\left(\gamma_i\,\frac{a_i c_i + \lambda\,\gamma_i e_i}{\alpha_i^2 + \lambda\,\gamma_i^2} - e_i\right)^2 + \sum_{i=r+1}^{p} e_i^2$$

$$f(\lambda) = \sum_{i=1}^{r}\left(\alpha_i\,\frac{\gamma_i c_i - d_i e_i}{\alpha_i^2 + \lambda\,\gamma_i^2}\right)^2 + \sum_{i=r+1}^{p} e_i^2\ . \qquad (2.22)$$

Obviously $f(A)$ is a rational function in A with sane poles. If $\alpha_i(\gamma_i c_i - \alpha_i e_i) \neq 0$ for some $1 \leq i \leq r$ then

$$\lambda = -\frac{\alpha_i^2}{\gamma_i^2} = -\mu_i$$

is a pole of f .

If $\underline{A}^T\underline{b} = \underline{0}$ and $\underline{C}^T\underline{d} = 0$ then from the definition of f (2.19), (2.18) we have

$$f(A) = \|\underline{d}\|^2 = \text{const.}$$

From (2.19) we see that $f(A) > 0$.

To prove the limit property in (4) we observe that the solution $x(\lambda)$ of (2.17a) converges for $\lambda \to \infty$ to a solution of $\underline{C}^T\underline{C}\underline{x} = \underline{C}^T\underline{d}$. Though $\underline{x}(\infty)$ may not be $\underline{C}^{+}\underline{d}$ it has the same residual. Therefore:

$$\|C\,x(\infty) - d\|^2 = \|C\,C^{+}d - d\|^2 \quad .$$

This property can of course also be seen from the representation (2.22).

To prove (5) we differentiate (2.19) and (2.17a)

$$f'(h) = 2x'(\lambda)^T C^T(C\,x(\lambda) - d) \tag{2.23}$$

$$(A^TA + \lambda C^TC)x'(\lambda) = -C^T(C\,x(h) - d) \quad . \tag{2.24}$$

Now since $\lambda > 0$ we can perform the Cholesky decomposition

$$(A^TA + \lambda C^TC) = R_\lambda^T R_\lambda \quad . \tag{2.25}$$

Using (2.25) and (2.24) in (2.23) gives

$$f'(A) = -2\|R_\lambda^{-T} C^T(C\,x(\lambda) - d)\|^2 < 0 \quad . \tag{2.26}$$

Every solution λ of the secular equation (2.19) with $A \neq -\mu_i$ defines $x(\lambda)$ (2.18) and $(x(\lambda),\lambda)$ is a solution of the normal equations (2.17a,b).

But there might be more solutions for some $A = -\mu_i$. In this case the matrix of (2.17a) is singular. If a solution should exist, the system must be consistent. This is especially true if $A^Tb = C^Td = 0$ then $x(\lambda)$ is an eigenvector to $\lambda = -\mu_i$. The next lemma describes the condition for consistency:

Lemma 2.2. Let μ_i be an eigenvalue of $\det(A^TA - \mu C^TC) = 0$. Then

$$(A^TA - \mu C^TC)x = A^Tb - \mu_i C^Td \tag{2.27}$$

is a consistent system if one of the conditions holds:

(i) $\lim_{A \to -\mu_i} f(A) = \lim_{A-1 \to -\mu_i} \|C\,x(\lambda) - d\|^2$ exists, i.e., $-\mu_i$ is not a pole of f .

(ii) Let $J = \{j \mid 1 \leq j \leq k\ ,\ \frac{\alpha_j^2}{\gamma_j^2} = \mu_i\}$ then $\alpha_j(c_j\gamma_j - \alpha_j e_j) = 0$

for $j \in J$, where all variables are defined in the proof of Lemma 2.1.

Proof. We prove first (ii) using the BSVD to transform (2.27).

$$A = UD_A X^{-1}\ ,\ D_A = \mathrm{diag}(\alpha_1,\ldots,\alpha_n)$$

$$C = V D_C X^{-1}\ ,\ D_C = \mathrm{diag}(\gamma_1,\ldots,\gamma_r,0\ldots0) \quad .$$

With $y := X^{-1}x$, $c := U^Tb$, $e := V^Td$, $(A^TA + A C^TC)x = A^Tb + \lambda C^Td$ becomes

$$\left.\begin{aligned} (\alpha_k^2 + \lambda\gamma_k^2)y_k &= \alpha_k c_k + \lambda\gamma_k e_k, \quad k = 1,\ldots,r \\ \alpha_k^2 y_k &= \alpha_k c_k, \quad k = r+1,\ldots,n \quad . \end{aligned}\right\} \tag{2.28}$$

Now observe that r = rank Of C and because $NS(A) \cap NS(C) = \{0\}$, $\alpha_k \neq 0$, $k = r+1,\ldots,n$. If $\lambda = -\mu_i = -\alpha_i^2/\gamma_i^2$, then for every equation j of (2.28) with

$$\alpha_j^2 - \mu_j\gamma_j^2 = 0$$

i.e., for all $j \in J = \{j \mid 1 \leq j \leq k\ ,\ \mu_i = \frac{\alpha_i^2}{\gamma_i^2} = \frac{\alpha_j^2}{\gamma_j^2}\}$, the right hand side must be zero:

$$a_{JJ}c_J - \frac{a_i^2}{\gamma_i^2}\gamma_J e_J = \alpha_j c_j - \frac{\alpha_j^2}{\gamma_j^2} e_j = 0$$

$$\Rightarrow \alpha_j(\gamma_j c_{JJ} - \gamma_j e_j) = 0 \quad \text{for } j \in J \quad .$$

We note that the solution of (2.28) is given by

$$y_k = \left\{ \begin{array}{ll} (\alpha_k c_k - \mu_i \gamma_k e_k) / (\alpha_k^2 - \mu_i \gamma_k^2), & k \notin J \\ \text{arbitrary}, & k \in J \\ c_k/\alpha_k, & k = r+1,\ldots,n \end{array} \right\} \qquad (2.29)$$

Furthermore

$$\tilde{y}_k := \lim_{\lambda \to -\mu_i} y_k(\lambda) = \left\{ \begin{array}{ll} (\alpha_k c_k - \mu_i \gamma_k e_k) / (\alpha_k^2 - \mu_i \gamma_k^2), & k \notin J \\ e_k/\gamma_k, & k \in J \\ c_k/\alpha_k, & k = r+1,\ldots,n \end{array} \right\} \qquad (2.30)$$

From (2.30) we conclude that

$$\lim_{A \to -\mu_i} x(\lambda) = \lim_{A \to -\mu_i} X y(\lambda) = X \tilde{y}$$

exists and therefore

$$\lim_{\lambda \to -\mu_i} \|C x(\lambda) - d\|^2$$

exists. This implies that f has no pole for $A = -\mu_i$.

Conversely if f has not a pole for $\lambda = -\mu_i$ then also

$$\lim_{A \to -\mu_i} y(\lambda)$$

must exist, which means that the system is consistent. □

We can now characterize the solutions of the normal equations precisely:

Theorem 2.4. Let f be the length function defined by (2.19),(2.18). If $f(A) = \alpha^2$ and $\det(A^T A + \lambda C^T C) \neq 0$ then there exists a unique $x(\lambda)$ so that $(x(\lambda),\lambda)$ solves the normal equations. If

$$\det(A^T A - \mu_i C^T C) = 0 \text{ and } \lim_{\lambda \to -\mu_i} f(A) \leq \alpha^2 \qquad (2.31)$$

then there exist a $x(-\mu_i)$ so that $(x(-\mu_i),-\mu_i)$ solves the normal equations, but $x(-\mu_i)$ is only unique if $\lim_{A \to -\mu_i} f(A) = \alpha^2$.

Proof. We have only to prove (2.31). We use the terminology defined in the proof of Lemma 2.2. For $A = -\mu_i$ the general solution y of the transformed normal equations (2.28) is given by (2.29). We have to see if we can satisfy the equation

$$\|D_c y - e\| = \alpha^2 \qquad (2.32)$$

with this solution y. In componends (2.32) is

$$\sum_{k \notin J} (\gamma_k y_k - e_k)^2 + \sum_{k \in J} (\gamma_k y_k - e_k)^2 + \sum_{k=r+1}^{n} e_k^2 = \alpha^2 . \qquad (2.33)$$

We can choose the componends of y_k, $k \in J$ arbitrarily. From (2.30) we see that

$$\lim_{\lambda \to -\mu_i} f(A) = \lim_{\lambda \to -\mu_u} \|D_c y(\lambda) - e\|^2 = \|D_c \tilde{y} - e\|^2$$

$$\lim_{\lambda \to -\mu_i} f(A) = \sum_{k \notin J} (\gamma_k y_k - e_k)^2 + \sum_{k=r+1}^{n} e_k^2 . \qquad (2.34)$$

Therefore $\|D_c y - e\|^2$ is minimized for $y = \tilde{y}$.

From (2.34) we see that if

$$\lim_{A \to -\mu_i} f(A) > \alpha^2$$

we can not determine a solution $\underline{y}(\lambda)$ that satisfies (2.32). If

$$\lim_{h \to -\mu_i} f(A) = \alpha^2$$

then the unique solution is $\tilde{y}$ (2.30). If

$$\lim_{A \to -\mu_i} f(A) < \alpha^2$$

then we can choose y_k, $k \in J$ so that

$$\sum_{k \in J} (\gamma_k y_k - e_k)^2 = \alpha^2 - \lim_{A \to -\mu_i} f(A) \qquad (2.35)$$

and $\underline{y}(-\mu_i)$ is not unique. □

We remark that the componends of $\underline{y}$: y_k, $k \in J$ are the (possibly) non-zero componends of the eigenvector $\underline{y}_{\mu_i}$ of

$$(\underline{D}_A^T \underline{D}_A - \mu_i \underline{D}_c^T \underline{D}_c)\underline{y}_{\mu_i} = \underline{0} .$$

Therefore a solution for the case where $\lim_{A \to -\mu_i} f(A) < \alpha^2$ can also be written in the following way. Take any eigenvector $\underline{y}_{\mu_i}$ belonging to μ_i and determine ρ such that

$$\|\underline{D}_c(\tilde{\underline{y}} + P \underline{y}_{\mu_i}) - \underline{d}\| = \alpha^2 .$$

Then $(\tilde{\underline{y}} + \rho \underline{y}_{\mu_i}, -\mu_i)$ is the solution of the normal equations.

2.3 The Solution for the Equality Constraint.

We now consider problem (PlE)

$$\|\underline{A}\underline{x} - \underline{b}\|^2 = \min$$

subject to $\|\underline{C}\underline{x} - \underline{d}\| = \alpha^2$.

We continue to assume that the nullspaces of $\underline{A}$ and $\underline{C}$ intersect trivially. To find the solution we first have to compute the rightmost solution A of the secular equation. If $A^* > 0$ then $\underline{x}(\lambda^*)$ is the unique solution of (PlE). If $A^* < 0$ then we have to compute the smallest eigevalue μ_r of the generalized eigenvalue problem

$$\det(\underline{A}^T\underline{A} - \mu \underline{C}^T\underline{C}) = 0$$

and an eigenvector $\underline{x}_r$. If $A^* > -\mu_r$ then again by Theorem 2.1 $\underline{x}(\lambda^*)$ is the unique solution. However if $A^* \leq -\mu_r$ then a solution has the form

$$\underline{x}(\rho) = \lim_{\lambda \to -\mu_r} \underline{x}(\lambda) + \rho \underline{x}_r$$

where we have to determine ρ such that

$$\|\underline{C}\ \underline{x}(\rho) - \underline{d}\|^2 = \alpha^2 .$$

If $\underline{A}^T\underline{b} = \underline{C}^T\underline{d} = \underline{0}$ then $f(A) \equiv \|\underline{d}\|^2$ and λ^* does not exist. Then $(\rho \cdot \underline{x}_r, -\mu_r)$ solves the problem if $\|\underline{d}\|^2 \leq \alpha^2$ where ρ has to be chosen so that $\|\rho\ \underline{C}\ \underline{x}_r - \underline{d}\| = \alpha^2$.

2.4 The Solution for the Inequality Constraint.

We are now ready to solve (P1):

$$\|\underline{A}\underline{x} - \underline{b}\|^2 = \min \tag{2.36}$$

subject to

$$\|\underline{C}\underline{x} - \underline{d}\|^2 \leq \alpha^2 \tag{2.37}$$

The weaker condition (2.37) simplifies the problem. Let $M = \{\underline{x} \mid \|\underline{A}\underline{x} - \underline{b}\| = \min\}$ denote the set of solutions of the unconstrained problem (2.36). If for some $x \in M$ we have $\|\underline{C}\underline{x} - \underline{d}\|^2 \leq \alpha^2$ then this x is a solution.

If $M \neq \{\underline{A}^+\underline{b}\}$ then $\mu_i = 0$ is an eigenvalue of $\det(\underline{A}^T\underline{A} - \mu\,\underline{C}^T\underline{C}) = 0$. Since 0 is never a pole of f (the normal equations (2.27) are consistent for $\lambda = 0$) we can define the number

$$\alpha^2_{max} := \lim_{\lambda \to 0} f(A) .$$

Let $\alpha^2_{min} := \|(\underline{C}\,\underline{C}^+ - \underline{I})\underline{d}\|^2 = \lim_{\lambda \to \infty} f(\lambda)$. Then (P1) has no solution if $\alpha < \alpha_{min}$. If $a \geq \alpha_{max}$ then

$$\underline{x}_0 = \lim_{\lambda \to 0} (\underline{A}^T\underline{A} + \lambda\,\underline{C}^T\underline{C})^{-1}(\underline{A}^T\underline{b} + \lambda\,\underline{C}^T\underline{d})$$

is the solution. Observe that $\underline{x}_0$ is in general not $\underline{A}^+\underline{b}$ [40]. Similarly if $\alpha = \alpha_{min}$ then we have to determine among the solutions x that minimize $\|\underline{C}\underline{x} - \underline{d}\|^2$ the solution that minimizes $\|\underline{A}\underline{x} - \underline{b}\|^2$ which is

$$\underline{x}_\infty = \lim_{\mu \to 0} (\mu\,\underline{A}^T\underline{A} + \underline{C}^T\underline{C})^{-1}(\mu\,\underline{A}^T\underline{b} + \underline{C}^T\underline{d}) .$$

In general we will have

$$a_{min} < a < a_{max} .$$

This means that we have to determine the unique positive solution λ^* of the secular equation

$$f(\lambda) = \alpha^2 .$$

And $\underline{x}(\lambda^*) = (\underline{A}^T\underline{A} + \lambda^*\,\underline{C}^T\underline{C})^{-1}(\underline{A}^T\underline{b} + \lambda^*\,\underline{C}^T\underline{d})$ will solve (P1).

The extreme cases $a = \alpha_{max}$ and $a = \alpha_{min}$ correspond to the two problems

$$\left.\begin{array}{l} \min \|\underline{C}\underline{x} - \underline{d}\| \\ \text{subject to} \\ \|\underline{A}\underline{x} - \underline{b}\| = \min \end{array}\right\} \quad \text{if } \alpha = \alpha_{max}$$

$$\left.\begin{array}{l} \min \|\underline{A}\underline{x} - \underline{b}\|^2 \\ \text{subject to} \\ \|\underline{C}\underline{x} - \underline{d}\| = \min \end{array}\right\} \quad \text{if } \alpha = \alpha_{min}$$

Both are of course only nontrivial if $\underline{A}$ respectively $\underline{C}$ is rank deficient.

To compute the solution in the case $\alpha_{min} < a < \alpha_{max}$ we proceed iteratively. The following algorithm shows the basic idea.

(a) start with $\lambda > 0$

(b) while not converged do

begin

solve the least square problem

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix} \underline{x}_\lambda \approx \begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\,\underline{d} \end{pmatrix}$$

correct A to solve $f(A) = \alpha^2$ where

$f(\lambda) = \|\underline{C}\underline{x}_\lambda - \underline{d}\|^2$ . .

end;

At every step of the iteration we have to solve a least squares problem. If we use an orthogonalization procedure [48] it is more expensive than solving the normal equations by the Cholesky decomposition [38,28,33]. Rut it is well known [26] that it is numerically preferable to solve the least square problem by orthogonal transformations.

We shall show in Section 5 how to reduce the amount of work using a structure of the matrices A and C or by using an orthogonal decomposition of them.

In the next two sections we shall consider the special cases $\underline{A} = \underline{I}$ and $\underline{C} = \underline{I}$. It is interesting that we can formulate dual equations for these special cases. Furthermore Eldén [8] has showed how to reduce the general problem to a problem with $\underline{C} = \underline{I}$.

3. The Relaxed Least Squares Problem.

We consider in this section the special case of a least squares problem with a quadratic constraint where $\underline{C} = \underline{I}$:

$$\begin{aligned} &\|\underline{\underline{A}}\underline{x} - \underline{b}\| = \min \\ &\text{subject to} \\ &\|\underline{x}\| = \alpha \quad . \end{aligned} \tag{P2E}$$

The problem with the inequality constraint $\|\underline{x}\| \leq \alpha$ (P2) was called by Rutishauser [35] the relaxed least squares problem. Problem (P2E) can be interpreted to find the minimum of a quadratic form $\|\underline{\underline{A}}\underline{x} - \underline{b}\|^2$ on the sphere $\|\underline{x}\|^2 = \alpha^2$, which is a special case of finding the stationary values of a quadratic form on a sphere. This problem has been analyzed by Forsythe and Golub [10] in 1965. The two authors proved that if $(\underline{x}_1, \lambda_1)$ and $(\underline{x}_2, \lambda_2)$ are two solutions of the normal equations then

$$\lambda_1 > \lambda_2 \Rightarrow \|\underline{\underline{A}}\underline{x}_2 - \underline{b}\|^2 > \|\underline{\underline{A}}\underline{x}_1 - \underline{b}\|^2 \ .$$

Their proof is rather complicated. Kahan [23] gave an elementary proof and stated the theorem

$$\|\underline{\underline{A}}\underline{x}_2 - \underline{b}\|^2 - \|\underline{\underline{A}}\underline{x}_1 - \underline{b}\|^2 = \frac{\lambda_1 - \lambda_2}{2} \|\underline{x}_2 - \underline{x}_1\| \tag{3.1}$$

which is Theorem 2.1 for $\underline{C} = \underline{I}$. Unfortunately this theorem was never published. In 1972 Spjøtvoll [39] gave a simpler proof than Forsythe - Golub but did not quite obtain Kahan's result. He showed that

$$\|\underline{\underline{A}}\underline{x}_2 - \underline{b}\|^2 - \|\underline{\underline{A}}\underline{x}_1 - \underline{b}\|^2 = (\lambda_1 - \lambda_2)(\alpha^2 - \underline{x}_1^T \underline{x}_2)$$

but did not see that $\alpha^2 = \frac{1}{2}(\|\underline{x}_1\|^2 + \|\underline{x}_2\|^2)$.

3.1 Results from the General Theory.

Considering the theory of Section 2 and putting $\underline{C} = \underline{I}$ and $\underline{d} = \underline{0}$ we get the normal equations

$$(\underline{A}^T\underline{A} + A\underline{I})\underline{x} = \underline{A}^T\underline{b} \tag{3.2}$$

$$\|\underline{x}\|^2 = \alpha^2 \ . \tag{3.3}$$

Theorem 2.1 gives us equation (3.1). This means that the solution of (P2E) is the solution $(\underline{x},\lambda)$ of (3.2), (3.3) with the largest A . Since $NS(\underline{A}) \cap NS(\underline{I}) = \{\underline{0}\}$ we have from Theorem 2.3 that the solution of (P2E) is unique if $A \neq -\sigma_i^2$ (where σ_i is a singular value of $\underline{A}$).

The length function f can be written

$$f(A) = \|(\underline{A}^T\underline{A} + A\underline{I})^{-1}\underline{A}^T\underline{b}\|^2 \ . \tag{3.4}$$

Using the singular value decomposition of $\underline{A}$ [14]

$$\underline{A} = \underline{U}\underline{\Sigma}\underline{V}^T$$

where

$\underline{U}, \underline{V}$ orthogonal

$\underline{\Sigma} = \mathrm{diag}(\sigma_1,\ldots,\sigma_p)$, $p = \min(m,n)$.

The right hand side of (3.4) becomes

$$f(\lambda) = \|(\underline{\Sigma}^+\underline{\Sigma} + \lambda\underline{I})^{-1}\underline{\Sigma}^T\underline{C}\| \quad \text{with } \underline{C} := \underline{U}^T\underline{b} \ .$$

In componends this is

$$f(\lambda) = \sum_{i=1}^{p}\left(\frac{\sigma_i c_i}{\sigma_i^2 + \lambda}\right)^2 \ . \tag{3.5}$$

If $\underline{A}^T\underline{b} = \underline{0}$ then $c = \underline{0}$ and $f(A) \equiv 0$. Theorem 2.4 and the fact that $\lim_{\lambda\to\infty} f(A) = 0$ implies that for every α we have a solution of (P2E) which is not unique if $A = -\sigma_i^2$ and $\lim_{A\to-\sigma_i^2} f(A) < \alpha^2$.

This was already stated by Forsythe and Golub in 1965 [10].

For the problem with the inequality constraint $\|\underline{x}\|^2 \leq \alpha^2$ (P2) we have either

$$\|\underline{A}^+\underline{b}\|^2 \leq \alpha^2$$

then the solution is $\underline{x} = \underline{A}^+\underline{b}$ or if $\|\underline{A}^+\underline{b}\|^2 > \alpha^2$ then we have to determine the unique solution A* of the secular equation

$$f(\lambda) = \alpha^2 \quad \text{in } (0,\infty) \ .$$

The solution is in that case $\underline{x}(\lambda^*)$. We note that we can reduce the length of $\underline{x}$ arbitrarily by choosing α small.

3.2 The Dual Normal Equations.

Theorem 3.1.

(i) Let $(\underline{x},\lambda)$ be a solution of the primal normal equation

$$(\underline{A}^T\underline{A} + A\underline{I})\underline{x} = \underline{A}^T\underline{b} \tag{3.6}$$

$$\|\underline{x}\|^2 = \alpha^2 \tag{3.7}$$

with $\lambda \neq 0$. Then $(\underline{z},\lambda)$ with

$$\underline{z} = \frac{1}{\lambda}(\underline{A}\underline{x} - \underline{b}) \tag{3.8}$$

is a solution of the dual normal equations

$$(AA^T + \lambda I)z = -b \tag{3.9}$$

$$\|A^T z\| = \alpha^2 \ . \tag{3.10}$$

(ii) Let (z,λ) be a solution of the dual normal equations (3.8), (3.9). Then (x,λ) with

$$x = -A^T z \tag{3.11}$$

is a solution of the prime normal equations (3.6), (3.7).

Proof.

(i) $$(AA^T + \lambda I)\frac{1}{\lambda}(Ax - b) = \frac{1}{\lambda}A(\underbrace{A^TA + \lambda I)x}_{A^Tb} - \frac{1}{\lambda}AA^Tb - b$$

$$= -b \ .$$

Furthermore

$$\|A^Tz\|^2 = \|A^T\frac{1}{\lambda}(Ax - b)\|^2 = \|x\|^2 = \alpha^2 \ .$$

(ii) $$(A^TA + \lambda I)(-A^Tz) = -A^T(\underbrace{AA^T + \lambda I)z}_{-b} \ .$$

And

$$\|-A^Tz\|^2 = \|x\|^2 = \alpha^2 \ . \ \square$$

Theorem 3.1 shows that we may simplify computations to solve (P2) or (P2E). Since A is an $(m \times n)$ matrix one of the linear systems (3.6) or (3.8) will be smaller and it may be more economical to iterate with the smaller.

But Theorem 3.1 also gives theoretical equations.

Corollary 3.1. Let A be an $(m \times n)$ matrix and $A \neq -\sigma_i^2$ where σ_i is a singular value of A. Then

$$(A^TA + \lambda I)^{-1}A^T = A^T(AA^T + A I)^{-1} \tag{3.12}$$

$$(AA^T + \lambda I)^{-1} = \frac{1}{\lambda}(I - A(A^TA + \lambda I)^{-1}A^T) \ . \tag{3.13}$$

Proof. From Theorem 3.1 (3.12) follows by equating x and (3.13) by equating z for the dual and primal equations. $\square$

As an application we consider (3.13) for $n = 1$ (i.e., A is a vector) and $\lambda = 1$:

$$(I + aa^T)^{-1} = I - \frac{1}{(a^Ta + 1)}aa^T \ . \tag{3.14}$$

We observe that Corollary 3.1 is a special case of the Shermann - Morrison-Woodbury formula:

Let A be $(n \times n)$, U, V (nxp) matrices, then every solution x of

$$(A + UV^T)x = b \ , \tag{3.15}$$

is also the solution of the augmented system

$$Ax + Uy = b \tag{3.16a}$$

$$V^Tx - y = 0 \ . \tag{3.16b}$$

Now assume that A is nonsingular. Then we have from (3.16a)

$$x = A^{-1}b - A^{-1}Uy \ .$$

Introducing this in (3.16b) we get

$$\underline{y} = (\underline{I} + \underline{V}^T \underline{A}^{-1} \underline{U})^{-1} \underline{V}^T \underline{A}^{-1} \underline{b}$$

$$\Rightarrow \underline{x} = \underline{A}^{-1}\underline{b} - \underline{A}^{-1} \underline{U}(\underline{I} + \underline{V}^T \underline{A}^{-1} \underline{U})^{-1} \underline{V}^T \underline{A}^{-1} \underline{b} .$$

But from (3.15) and (3.16b) we have also

$$x = (\underline{A} + \underline{U}\,\underline{V}^T)^{-1} b$$

$$Y = \underline{V}^T(\underline{A} + \underline{U}\,\underline{V}^T)^{-1} b .$$

Equating both expressions for $\underline{x}$ and $\underline{y}$ we get the matrix equations

$$\underline{V}^T(\underline{A} + \underline{U}\,\underline{V}^T)^{-1} = (\underline{I} + \underline{V}^T \underline{A}^{-1} \underline{U})^{-1} \underline{V}^T \underline{A}^{-1}$$

$$(\underline{A} + \underline{U}\,\underline{V}^T)^{-1} = \underline{A}^{-1}(\underline{I} - \underline{U}(\underline{I} + \underline{V}^T \underline{A}^{-1} \underline{U})^{-1} \underline{V}^T \underline{A}^{-1}) .$$

(Shermann - Morrison - Woodbury formulas)

we note that for $A := \lambda I$ and $\underline{U} = \underline{V} := \underline{A}$ we get the equations (3.12) and (3.13).

We finally remark that if $A > 0$ the dual equations are the normal equations of the least squares problem

$$\begin{pmatrix} \underline{A}^T \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix} \underline{z} \approx \begin{pmatrix} \underline{0} \\ -\frac{1}{\sqrt{\lambda}}\,\underline{b} \end{pmatrix} .$$

The dual equations have no solution if $A = 0$. We see this clearly because of the factor $-\frac{1}{\sqrt{\lambda}}$, but also from (3.9) since it is clear that for $\lambda = 0$ a solution $\underline{z}$ of (3.9) may not exist for arbitrary b and $m > n$.

3.3 Eldén's Transformation.

Eldén [8] has shown how to transform a problem (P1) to a problem (P2) using orthogonal transformations. To illustrate his idea consider the least square representation of (P1)

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\,\underline{d} \end{pmatrix} \tag{3.17}$$

$$\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2 . \tag{3.18}$$

We may assume that the rank of $C = p < n$. (C is a $(p \times n)$ matrix.) If not then we can perform a QR decomposition with column pivoting

$$\underline{C} = \underline{Q}\begin{pmatrix} \underline{R} \\ \underline{0} \end{pmatrix}\underline{P}^T , \quad \underline{R} = \begin{pmatrix} 0 & \square \end{pmatrix} .$$

Observe that $\|\underline{C}\underline{x} - \underline{d}\| = \|\underline{R}\,\underline{P}^T\underline{x} - \underline{Q}^T\underline{d}\|$. Therefore we can replace C by $\underline{R}$, $\underline{d}$ by $\underline{Q}^T\underline{d}$, $\underline{A}$ by $\underline{A}\underline{P}$, and $\underline{x}$ by $\underline{y} = \underline{P}^T\underline{x}$ in (3.17), (3.18). The problem for $\underline{y}$ has now a matrix $\underline{C}$ $(p \times n)$ with rank of $\underline{C} = p$. After it is solved we obtain $\underline{x} = \underline{P}\underline{y}$. The same preprocessing we can apply to $\underline{A}$ and $\underline{b}$ and so assume that $\mathrm{rank}(\underline{A}) = m \leq n$.

If $p = n$ then we can make the change of variables

$$x' := \underline{C}\underline{x} , \quad A' := \underline{A}\underline{C}^{-1}$$

which transforms the problem to

$$\begin{pmatrix} \underline{A}' \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix} \underline{x}' \approx \begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\,\underline{d} \end{pmatrix}$$

$$\|\underline{x}' - \underline{d}\|^2 = \alpha^2 .$$

We assume now $p < n$ and C has rank p. Then we make a QR decomposition of C^T:

$$C^T = (V_1, V_2) \begin{matrix} \underline{R} \\ \underline{0} \end{matrix} \begin{matrix} \}p \\ \}n-p \end{matrix} \tag{3.19}$$

with nonsingular triangular matrix R.

We change variables:

$$\underline{X} =: \underline{V}_1\underline{y}_1 + \underline{V}_2\underline{y}_2$$

and (3.17) becomes:

$$\begin{pmatrix} \underline{A}\underline{V}_1 & \underline{A}\underline{V}_2 \\ \sqrt{\lambda}\,\underline{R}^T & \underline{0} \end{pmatrix} \begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} \approx \begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\,\underline{V}_2^T\,\underline{d} \end{pmatrix}. \tag{3.21}$$

Whereas (3.18) is now

$$\|\underline{R}^T\,\underline{V}_1\,\underline{y}_1 - \underline{d}\|^2 = \alpha^2 \quad . \tag{3.22}$$

Now we perform another QR decomposition

$$\underline{A}\,\underline{V}_2 = (\underline{Q}_1, \underline{Q}_2) \begin{pmatrix} \underline{U} \\ \underline{0} \end{pmatrix} \} \begin{matrix} p \\ p-n \end{matrix} \tag{3.23}$$

Since we assumed that A has rank m and $\underline{V}_2$ is orthogonal, U is a non-singular upper triangular matrix. Now (3.21) is the same problem as:

$$\begin{pmatrix} \underline{Q}_1^T\,\underline{A}\,\underline{V}_1 & , & \underline{U} \\ \underline{Q}_2^T\,\underline{A}\,\underline{V}_1 & , & \underline{0} \\ \sqrt{\lambda}\,\underline{R}^T & , & \underline{0} \end{pmatrix} \begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} \approx \begin{pmatrix} \underline{Q}_1^T\,\underline{b} \\ \underline{Q}_2^T\,\underline{b} \\ \sqrt{\lambda}\,\underline{V}_2^T\,\underline{d} \end{pmatrix} \quad . \tag{3.23}$$

Now for every $\underline{y}_1$ we can determine $\underline{y}_2$ such that the first p-n equations are exactly satisfied. Therefore the problem splits as follows

$$\begin{pmatrix} \underline{Q}_2^T\,\underline{A}\,\underline{V}_1 \\ \sqrt{\lambda}\,\underline{R}^T \end{pmatrix} \underline{y}_1 \approx \begin{pmatrix} \underline{Q}_2^T\,\underline{b} \\ \sqrt{\lambda}\,\underline{V}_2^T\,\underline{d} \end{pmatrix} \tag{3.25}$$

with (3.22) as constraint. The final change of variables

$$\underline{y}_1' := \underline{R}^T\,\underline{y}_1$$

gives the desired problem (P2).

3.4 Rutishauser's Relaxed and Doubly Relaxed Least Squares Problem.

In [35] Rutishauser remarked that if we minimize

$$Q(\underline{x}) = \|\underline{A}\underline{x} - \underline{b}\|^2 \tag{3.26}$$

for a ill-conditioned matrix $\underline{A}$ (i.e., $\mathcal{K} = \|\underline{A}\|\|\underline{A}^+\| = \sigma_1/\sigma_r \gg 1$, where $\sigma_1 > \ldots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$ are the singular values of A), that then the exact solution $\hat{\underline{x}} = \underline{A}^+\underline{b}$ may be not satisfactory because $\|\hat{\underline{x}}\| \gg 1$ and evaluation of $\underline{A}\hat{\underline{x}}$ will be affected by cancellation.

Replacing (3.26) by

$$Q_\varepsilon(\underline{x}) = \|\underline{A}\underline{x} - \underline{b}\|^2 + \varepsilon^2\|\underline{x}\| \tag{3.27}$$

will give a shorter solution $\underline{x}_\varepsilon$ which however does not minimize (3.27) any more. The process of relaxing can be described as follows: The solution of (3.26) is the solution of the least squares problem

$$\underline{A}\underline{x} \approx \underline{b} \quad . \tag{3.28}$$

The relaxed solution $\underline{x}_\varepsilon$ of (3.27) is obtained by choosing some $\varepsilon > 0$ and solving

$$\begin{pmatrix} \underline{A} \\ \varepsilon \underline{I} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{b} \\ \underline{0} \end{pmatrix} \quad . \tag{3.29}$$

The connection to problem (P2) is as follows: Instead of presenting a bound for the length of x : $\|\underline{x}\|^2 \leq \alpha^2$ and solve the equation

$$f(\lambda) = \alpha^2$$

to determine $\lambda > 0$, we simply set $\lambda = \varepsilon^2 > 0$ and solve for x .

Rutishauser [35] defines the double relaxed solution $\underline{x}_{d\varepsilon}$ to be the solution of

$$(\underline{A}^T\underline{A} + \varepsilon^2 \underline{I} + \varepsilon(\underline{A}^T\underline{A} + \varepsilon^2 \underline{I})^{-1}]\underline{x} = \underline{A}^T\underline{b} \quad . \tag{3.30}$$

Observe that $\underline{x}_{d\varepsilon}$ is obtained by relaxing the normal equations of the relaxed solution:

$$\begin{pmatrix} \underline{A}^T\underline{A} + \varepsilon^2 \underline{I} \\ \varepsilon \underline{I} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{A}^T\underline{b} \\ \underline{0} \end{pmatrix} \quad . \tag{3.31}$$

If we put $\underline{B} := \underline{A}^T\underline{A} + \varepsilon^2 \underline{I}$ then the normal equations of (3.31) are

$$(\underline{B}^T\underline{B} + \varepsilon^2 \underline{I})x = \underline{B}^T \underline{A}^T \underline{b} \quad . \tag{3.32}$$

But $\underline{B}$ is symmetric and non-singular, therefore we may multiply 6.32) from left by $\underline{B}^{-1}$ and we get

$$(\underline{B} + \varepsilon^2 \underline{B}^{-1})\underline{x} = \underline{A}^T\underline{b}$$

which is (3.30). From (3.30) we have

Lemma 3.1. The double relaxed solution $\underline{x}_{d\varepsilon}$ minimizes

$$Q_d(\underline{x}) = \|(\underline{A}^T\underline{A} + \varepsilon\underline{I})\underline{x} - \underline{A}^T\underline{b}\|^2 + \varepsilon^2\|\underline{x}\|^2 \, .$$

The aim of relaxing is to approximate the "ideal solution" [13] without explicitely computing the singular value decomposition. By the "ideal" solution we mean the following: Let

$$\underline{A} = \underline{U}\,\underline{\Sigma}\,\underline{V}^T \, , \; \underline{U} \, , \; \underline{V} \text{ orthogonal}$$

$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \, , \quad \sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$$

be the singular value decomposition of A . If the datas and the computation were exact then the shortest solution of

$$\|\underline{A}\underline{x} - \underline{b}\|^2 = \min$$

would be $\underline{x} = \underline{A}^+\underline{b} = \underline{U}\underline{y}$ where y is defined by

$$y_i = \begin{cases} \dfrac{c_i}{\sigma_i} & \text{if } \sigma_i \neq 0 \qquad (\underline{c} := \underline{U}^T\underline{b}) \\ 0 & \text{if } \sigma_i = 0 \, . \end{cases}$$

In practical computation however it is not clear which σ_i are to be interpreted as zero [11,26]. One is forced to make a rank decision, to prescribe a tolerance τ and to compute

$$\tilde{y}_i = \begin{cases} \dfrac{c_i}{\sigma_i} & \text{if } \sigma_i \geq \tau \\ 0 & \text{if } \sigma_i < \tau \end{cases} \quad . \tag{3.33}$$

We will call $\tilde{y}_i$ the "ideal solution".

Now consider the function defined on $[0,\infty)$

$$k(u) := \begin{cases} 0 & \text{if } \sigma < \tau \\ \frac{1}{\sigma} & \text{if } \sigma \geq \tau \end{cases} \quad . \tag{3.34}$$

The coefficient of $\tilde{y}_i$ multiplying c_i is given by $k(\sigma_i)$. The length of y depends on the choice of τ and is

$$\|\tilde{\underline{y}}\|^2 = \sum_{\sigma_i \geq \tau} \left(\frac{c_i}{\sigma_i} \right)^2 \quad .$$

Let $\gamma := \|\tilde{\underline{y}}\| \; / \; \|\underline{b}\|$. We can ask the question: how to choose ε such that the relaxed and the doubly relaxed solution have the same or at least approximately the same length as the ideal solution. The answer is given in [13]. We have

$$\text{if } \varepsilon \geq \frac{1}{2\gamma} \Rightarrow \|\underline{x}_\varepsilon\| < \gamma\|\underline{b}\| = \|\tilde{\underline{y}}\|$$

and

$$\text{if } \varepsilon \geq \frac{1}{2\gamma^2} \Rightarrow \|\underline{x}_{d_\varepsilon}\| < \gamma\|\underline{b}\| = \|\tilde{\underline{y}}\| \quad .$$

If we transform the relaxed problem (3.29) and the doubly relaxed problem (3.31) using SVD we get

$$(\underline{y}_\varepsilon)_i = \frac{\sigma_i}{\sigma_i^2 + \varepsilon^2} \, c_i \tag{3.35}$$

$$(\underline{y}_{d_\varepsilon})_i = \frac{\sigma_i}{\sigma_i^2 + \varepsilon^2 + \frac{\varepsilon^2}{\sigma_i^2 + \varepsilon^2}} \, c_i \quad . \tag{3.36}$$

We see that we can think of (3.35), (3.36) as approximations of (3.33). More precisely, we want to choose ε so that the two functions

$$k_\varepsilon(\sigma) := \frac{\sigma}{\sigma^2 + \varepsilon^2} \tag{3.37}$$

and

$$k_{d_\varepsilon}(\sigma) := \frac{\sigma}{\sigma^2 + \varepsilon^2 + \frac{\varepsilon^2}{\sigma^2 + \varepsilon^2}} \tag{3.38}$$

approximate the function k(o) (3.34). Rutishauser [35] and Molinari [13] show that indeed k_{d_ε} is a better approximation than k_ε . More research could be done in this direction, e.g. relaxing the normal equations of the original problem

$$\begin{pmatrix} \underline{A}^T\underline{A} \\ \varepsilon\underline{I} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{A}^T\underline{b} \\ \underline{0} \end{pmatrix}$$

would yield the function

$$k_{1\varepsilon} := \frac{\sigma}{\sigma^2 + \frac{\varepsilon^2}{\sigma^2}}$$

which is somehow between both discussed above. It is clear that those different relaxed solutions have to be computed without forming the normal equations. A program for $\underline{x}_\varepsilon$ and $\underline{x}_{d_\varepsilon}$ is given in [13].

<u>Theorem 3.2.</u> Let $\underline{A}_\varepsilon = \begin{pmatrix} \underline{A} \\ \varepsilon\underline{I} \end{pmatrix}$ and $\underline{A}_\varepsilon^+ = (\underline{B}_\varepsilon^+, \underline{C}_\varepsilon)$. ($\underline{B}_\varepsilon^+$ is an $n \times n$ matrix.) Then for ε sufficiently small we have

$$\underline{B}_\varepsilon^+ = \underline{A}^+ + \sum_{j=1}^{\infty} (-1)^j \, \underline{A}^+ (\underline{A}^{+T} \, \underline{A}^+)^{2j} \, \varepsilon^{2j} \, ,$$

Proof. Let $A = U \Sigma V^T$ with $\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_r \\ 0 & & 0 \end{pmatrix}$ be the singular value decomposition, with $\sigma_1 \geq \ldots \geq \sigma_r > 0$. Then

$$A_\varepsilon^+ = (A^T A + \varepsilon^2 I)^{-1} A^T = (V \Sigma^T \underbrace{U^T U}_{I_m} \Sigma V^T + \varepsilon^2 I)^{-1} (V \Sigma^T U^T, I_\varepsilon)$$

$$= V(\Sigma^T \Sigma + \varepsilon^2 I)^{-1} (\Sigma^T U^T, \varepsilon V^T)$$

$$= V([(\Sigma^T \Sigma + \varepsilon^2 I)^{-1} \Sigma^T] U^T, \varepsilon(\Sigma^T \Sigma + \varepsilon^2 I)^{-1} V^T)$$

$$= [B_\varepsilon^+, C_\varepsilon]$$

$$B_\varepsilon^+ = V(\Sigma^T \Sigma + \varepsilon^2 I)^{-1} \Sigma^T U^T$$

$$B_\varepsilon^+ = V \begin{vmatrix} \frac{\sigma_1}{a_1^2 + \varepsilon^2} & & 0 \\ & \ddots & \\ & \frac{\sigma_k}{\sigma_r^2 + \varepsilon^2} & \\ 0 & & 0 \end{vmatrix} U^T$$

now $\dfrac{a_i}{a_i^2 + \varepsilon^2} = \dfrac{1}{\sigma_i} \dfrac{1}{1 + (\varepsilon/\sigma_i)^2} = \dfrac{1}{\sigma_i} \sum_{j=0}^{\infty} (\varepsilon/\sigma_i)^{2j}$ for $|\varepsilon| < \sigma_r$,

$$B_\varepsilon^+ = \sum_{j=0}^{\infty} V \underbrace{\begin{bmatrix} 1/\sigma_1^{2j+1} & & 0 \\ & \ddots & \\ & 1/\sigma_r^{2j+1} & \\ 0 & & 0 \end{bmatrix}}_{\Sigma^+ (\Sigma^{+T} \Sigma^+)^{2j}} U^T \varepsilon^{2j}$$

where $\Sigma^+ = \begin{bmatrix} 1/\sigma_1 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_r \end{bmatrix}_{n}^{m}$,

we observe

$$V \Sigma^+ (\Sigma^{+T} \Sigma^+)^{2j} U^T = A^+ (A^{+T} A^+)^{2j} \quad . \quad 0$$

Remark. Theorem 3.2 shows that we can extrapolate $x_M = A^+ b$ from the relaxed solution $x_\varepsilon = B_\varepsilon^+ b$. Using $\varepsilon_i = \varepsilon_0 / 2^l$ we can apply the Romberg extrapolation since only even powers of ε occur. It has the advantage that the rank of A must not be determined.

. . .

4. Minimum Norm Solution with Given Norm of the Residual.

In this section we discuss the special case of a least squares problem with a quadratic constraint where $\underline{A} = \underline{I}$:

$$\left.\begin{aligned} \|\underline{x}\| &= \min \\ \text{subject to} \\ \|\underline{C}\underline{x} - \underline{d}\| &= \alpha \quad . \end{aligned}\right\} \qquad \text{(P3E)}$$

The problem with inequality constraint $\|\underline{C}\underline{x} - \underline{d}\| \leq \alpha$ will be denoted by (P3). Geometrically (P3E) means that we are looking for a point on the ellipsoid $\|\underline{C}\underline{x} - \underline{d}\|^2 = \alpha^2$ that is nearest to the origin. From the theory in Section 2 we have that the solution of (P3E) is a solution $(\underline{x},\lambda)$ of the normal equations

$$(\underline{I} + \lambda\, \underline{C}^T\underline{C})\underline{x} = \lambda\, \underline{C}^T\underline{d} \qquad (4.1)$$

$$\|\underline{C}\underline{x} - \underline{d}\| = \alpha^2 \qquad (4.2)$$

with largest λ . Since $NS(I) \cap NS(C) = \{0\}$ the solution is unique if $\lambda \neq -\gamma_i^2$ where γ_i is a singular value of C . The solution exists only if

$$\alpha \geq \|(\underline{C}\,\underline{C}^+ - \underline{I})\underline{d}\| := \alpha_{min} \quad .$$

For problem (P3) the solution is $\underline{x} = \underline{0}$ if $\alpha \geq \|\underline{d}\|$. The length function f is in this case

$$f(h) = \|(\lambda\underline{C}(\underline{I} + \lambda\, \underline{C}^T\underline{C})^{-1}\, \underline{C}^T - \underline{I})\underline{d}\|^2 \quad . \qquad (4.3)$$

By Corollary 3.1, (3.13), this is equal to

$$f(h) = \|(\underline{I} + \lambda\, \underline{C}\underline{C}^T)^{-1}\, \underline{d}\|^2 \qquad (4.5)$$

which gives again the connection to the dual equations.

4.1 The Dual Normal Equations.

We transform the normal equations (4.1), (4.2) as follows: We multiply (4.1) from left by C giving

$$(\underline{I} + \lambda\, \underline{C}\underline{C}^T)\underline{C}\underline{x} = \lambda\, \underline{C}\underline{C}^T\, \underline{d} \quad . \qquad (4.6)$$

If we subtract from (4.6) on both sides $\lambda\, \underline{C}\underline{C}^T\, \underline{d} + \underline{d}$ we get

$$(\underline{I} + h\underline{C}\underline{C}^T)(\underline{C}\underline{x} - \underline{d}) = -\underline{d} \quad .$$

Finally introducing the new variable $\underline{z} = \underline{C}\underline{x} - \underline{d}$ we get the theorem:

Theorem 4.1.

(i) Let $(\underline{x},\lambda)$ be a solution of the primal normal equations

$$\begin{aligned} (\underline{I} + \lambda\, \underline{C}^T\underline{C})\underline{x} &= \lambda\, \underline{C}^T\, \underline{d} \\ \|\underline{C}\underline{x} - \underline{d}\|^2 &= \alpha^2 \end{aligned} \qquad (4.7)$$

then $(\underline{z},\lambda)$ with $\underline{z} := \underline{C}\underline{x} - \underline{d}$ is a solution of the dual normal equations

$$\begin{aligned} (\underline{I} + \lambda\, \underline{C}\underline{C}^T)\underline{z} &= -\underline{d} \\ \|\underline{z}\|^2 &= \alpha^2 \quad . \end{aligned} \qquad (4.8)$$

(ii) Let $(\underline{z},\lambda)$ be a solution of (4.8), then $(\underline{x},\lambda)$ with $x = -h\, \underline{C}^T\, z$ is a solution of (4.7).

Proof.

(i) $$\begin{aligned} (\underline{I} + \lambda\, \underline{C}\underline{C}^T)(\underline{C}\underline{x} - \underline{d}) &= \underline{C}(\underline{I} + \lambda\, \underline{C}^T\underline{C})\underline{x} - \underline{d} - \lambda\, \underline{C}\underline{C}^T\, \underline{d} \\ &= \underline{C}(\lambda\, \underline{C}^T\, \underline{d}) - \underline{d} - \lambda\, \underline{C}\underline{C}^T\, \underline{d} = -\underline{d} \quad . \end{aligned}$$

and

$$\alpha^2 = \|z\|^2 = \|Cx - d\|^2 .$$

(ii) $$(I + h\, C^T C)(-\lambda\, C^T z) = -h\, C^T(I + h\, CC^T)z = h\, C^T b .$$

$$\|Cx - d\|^2 = \|C(-\lambda\, C^T z) - d\|^2 = \|z\|^2 = \alpha^2 . \quad \square$$

Equating both expressions for x and z again gives us the identities of Corollary 3.1.

In contrast to the relaxed least squares problem, the dual equations exists for every λ .

4.2 Representation as Least Squares Problem.

If we want to solve (P3) then the solution is $x = 0$ if $\alpha > \|d\|$ and exists only if $\alpha > \alpha_{min} = \|(C C^+ - I)d\|$. For

$$\alpha_{min} < \alpha < \|d\| .$$

We have to compute the solution iteratively solving the secular equation $f(h) = \alpha^2$ where

$$f(\lambda) = \|Cx - d\|^2 = \|z\|^2 .$$

x and z are obtained solving either

$$\begin{pmatrix} I \\ \sqrt{\lambda}\, C \end{pmatrix} x \approx \begin{pmatrix} 0 \\ \sqrt{\lambda}\, d \end{pmatrix}$$

or the dual problem

$$\begin{pmatrix} I \\ \sqrt{\lambda}\, C^T \end{pmatrix} z \approx \begin{pmatrix} -d \\ 0 \end{pmatrix}$$

5. Computational Aspects.

To compute the solution of a least squares problem with a quadratic constraint we have to determine iteratively the largest solution of the secular equation. For the problem (PIE), (P2E) and (P3E) we can expect numerical difficulties since at every step of the iteration we have to solve a system of equations with the matrix

$$A^T A + \lambda\, C^T C \quad \text{where } h < 0 \tag{5.1}$$

which is in general not positive definite. A good way to solve this system is to transform it to diagonal form using BSVD of von Loan [45]. If $A = I$ or $C = I$ we can use SVD to diagonalize the system stably [48]. To avoid the forming of $A^T A$ and $C^T C$ we could consider the augmented $(m+n+p) \times (m+n+p)$ system

$$\begin{pmatrix} -1 & A & 0 \\ A^T & 0 & C^T \\ 0 & C & \frac{1}{\lambda} I \end{pmatrix} \begin{pmatrix} y \\ x \\ z \end{pmatrix} = \begin{pmatrix} b \\ 0 \\ d \end{pmatrix} . \tag{5.2}$$

Björk has shown [1] that for the augmented system

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} \tag{5.3}$$

We can obtain the condition number $2 \cdot \kappa(A)$ with appropriate scaling (instead of $\kappa^2(A)$ for the normal equations $A^T A x = A^T b$). However in (5.2) the condition number depends on the value of λ and may be large if h is near $-\mu_i$ where μ_i is an eigenvalue of $\det(A^T A - \mu\, C^T C) = 0$.

The solution is numerically much better defined for problem (P1), (P2) and (P3) where $\lambda > 0$. We can formulate the normal equations as least squares problems and besides (BSVD)- and SVD-diagonalization, there are several different ways to compute stably and cheaply the solution, especially if we can use a structure of the matrix. Since we are solving the secular equation with some iterative method we may have to compute derivatives of the length function f (see Section 5.4). If the iterative method converges fast it may not be necessary to transform the problem at all (especially if $\underline{A}$ and $\underline{C}$ are sparse). In general, however, it appears to be best [8], [12] to bidiagonalize the matrix for problems (P2) and (P3) before iterating.

5.1 Solution of a Relaxed Least Squares Problem with Band Matrix.

We consider (P2) with $\underline{A}$ respectively (P3) with $\underline{C}$ being a band matrix. For a given $\lambda > 0$ we have to solve the least squares problems

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{b} \\ \underline{0} \end{pmatrix} \tag{5.4}$$

respectively

$$\begin{pmatrix} \underline{I} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{0} \\ \sqrt{\lambda}\,\underline{d} \end{pmatrix} \quad . \tag{5.5}$$

If we interchange the equations in (5.5) we have in either case a least squares problem of the form

$$\begin{pmatrix} \underline{B} \\ \underline{D} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{c}_1 \\ \underline{c}_2 \end{pmatrix} \tag{5.6}$$

where $\underline{D}$ is a diagonal matrix ($\underline{I}$ or $\sqrt{\lambda}\,\underline{I}$) and $\underline{B}$ is a band matrix ($\underline{A}$ or $\sqrt{\lambda}\,\underline{C}$).

Using Givens-rotations we shall show how an orthogonal matrix G can be found such that

$$\underline{G}\begin{pmatrix} \underline{B} \\ \underline{D} \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{F} \end{pmatrix} \tag{5.7}$$

where $\underline{F}$ is a band upper triangular matrix. If the same rotations are applied to the right hand side the solution can be found by backsubstitution using $\underline{F}$.

To describe the transformation we assume that B is an $(m \times n)$ matrix with

$$b_{ij} = 0 \quad \text{if } i-j > m_1 \text{ or } j-i > m_2 \ .$$

This means that $\underline{B}$ has m_1 diagonals below the main diagonal and m_2 diagonals above, that contains the possibly non-zero elements.

Let

$$k_{min}(k) := \max\{k-m_2, 1\}$$

$$k_{max}(k) := \min\{k+m_1, m\} \ .$$

Then the k-th column of B has the only (possibly) non-zero elements

$$b_{ik} \ , \quad i = k_{min}(k), \ldots, k_{max}(k) \ .$$

In n-steps we now perform the transformation (5.7) annihilating in the k-th step the k-th column of B and producing the k-th row of F . Let $\underline{F}$ be at the beginning the diagonal matrix D and defined by

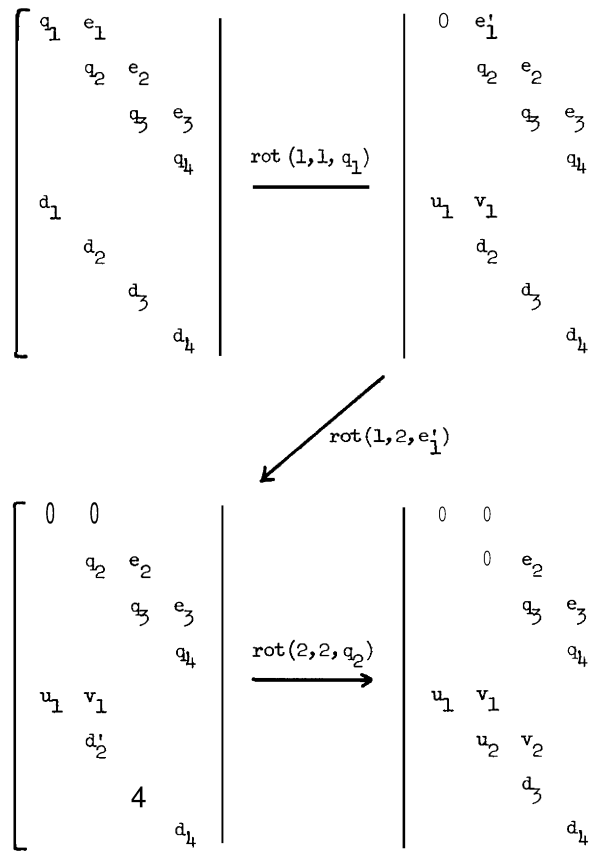
$$\text{rot}(i,k,x)$$

the multiplication of $\begin{pmatrix} \underline{B} \\ \underline{F} \end{pmatrix}$ from left by a Givens matrix that changes the **i -th** row of $\underline{B}$ and the k-th row of F annihilating the element x . **Then** the transformation (5.7) is described by

```
for k := 1 step 1 until n do
for i := k_min(k) step 1 until k_max(k) do
    rot(i,k,b_ik);
```

In the k-th step of this algorithm we annihilate m_1+m_2+1 elements (or less at the border of $\underline{B}$) using $\sim 2(m_1+m_2+1)^2$ multiplications. For $i = k_{min} = k-m2$ the Givens rotation changes only 2 elements: b_{ik} and f_{kk} . **Then for** $i = k_{min}+1$, 4 elements are affected, etc., until for $i = k_{max} = k+m_1$ when $b_{k+m_1 k}$ is rotated to zero, we have to change m_1+m_2+1 elements. The whole transformation requires therefore $\sim 2n(m_1+m_2+1)^2$ multiplications which is comparable to Gaussian elimination of the normal equations.

Especially if $\underline{B}$ is upper bidiagonal the first step to annihilate the first and second column is done as follows:



The following procedure performs the transformation

$$\underline{G}\begin{pmatrix} \underline{B} \\ \underline{D} \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{F} \end{pmatrix}$$

where

$$\underline{B} = \begin{pmatrix} q_1 & e_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & e_{n-1} \\ 0 & & & q_n \end{pmatrix} \qquad \underline{F} = \begin{pmatrix} u_1 & v_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & v_{n-1} \\ 0 & & & u_n \end{pmatrix}$$

and. stores the Givens rotations in two arrays c0[i] , si[i

```
procedure updec (n,q,e,d,u,v,c0,si);
value n; integer n;
array q,e,d,u,v,c0,si;
begin integer i; real t,c,s,h;
     procedure rot(a,b);
     value a,b; real a,b;
     begin real t;
          if b = 0 then begin c := 0; s := 1 end
          else
               begin t := -a/b; c := 1/sqrt(1+t↑2); s := t*c end
     end;
     for i := 1 step 1 until n do u[i] := d[i];
     for i := 1 step 1 until n do
     begin
          rot(q[i],u[i]);
          c0[2*i-1] := c; si[2*i-1] := s;
          u[i] := -q[i]*s + u[i]*c;
          if i < n then
          begin
               v[i] := -e[i]*s; h := e[i]*c;
               rot(h,u[i+1]);
               c0[2*i] := c; si[2*i] := s;
               u[i+1] := -h*s + u[i+1]*c
          end if;
     end i;
end updec;
```


The following procedure performs the transformation

$$\underline{G}\begin{pmatrix} \underline{c}1 \\ \underline{c}1 \end{pmatrix}$$

and solves the bidiagonal system by backsubstitution.

```
procedure solve (n,u,v,c0,si,cl,c2,y);
value n; integer n;
array c0, si, cl, c2, y;
begin
     integer i; real t,c,s,h;
     for i := 1 step 1 until n do
     begin
          c := c0[2*i-1]; s := si[2*i-1];
          h := cl[i]*c + c2[i]*s;
          c2[i] := -cl[i]*s + c2[i]*c;
          cl[i] := h;
          if i < n then
          begin
               c := c0[2*i]; s := si[2*i];
               h := cl[i]*c + c2[i]*s;
               c2[i+1] := -cl[i]*s + c2[i+1]*c;
               cl[i] := h;
          end if;
     end i;
     y[nl := c2[n] / u[n];
     for i := n-1 step -1 until 1 do
     y[il := (c2[i] - v[i]*y[i+1]) / u[i];
end solve;
```

Alternatively we can avoid storing the rotations observing that if

$$\begin{pmatrix} B \\ D \end{pmatrix} y \approx \begin{pmatrix} c1 \\ c2 \end{pmatrix}$$

and if

$$G\begin{pmatrix} B \\ D \end{pmatrix} = \begin{pmatrix} 0 \\ F \end{pmatrix} \quad , \; G \text{ orthogonal}$$

then the normal equations are

$$F^T F y = B^T c1 + D c2 \quad .$$

Therefore we can also find y solving

$$F^T w = B^T c1 + D c2$$

$$F y = w \quad .$$

The new procedure updec does not contain the variable c0 and si . Otherwise it is the same as before. The procedure **solve** however changes much:

```
procedure solve2 (n,u,v,c,y);
value n; integer n;
array u,v,c,y;
comment solves F^T F y = c;
begin integer i
    y[1] := c[1]/u[1];
    for i := 2 step 1 until n do
        y[i] := (c[i] - v[i-1]*y[i-1])/u[i];
        y[n] := y[n]/u[n];
        for i := n-1 step -1 until 1 do
            y[i] := (y[i] - v[i]*y[i+1])/u[i];
end solve2;
```

Before calling solve2 the right hand side c has to be defined by

$$c := B^T c1 + D c2 \quad .$$

5.2 Solution of a Least Squares Problem with Two Band Matrices.

The algorithm given in Section 5.1 can be generalized to solve a least square problem of the type

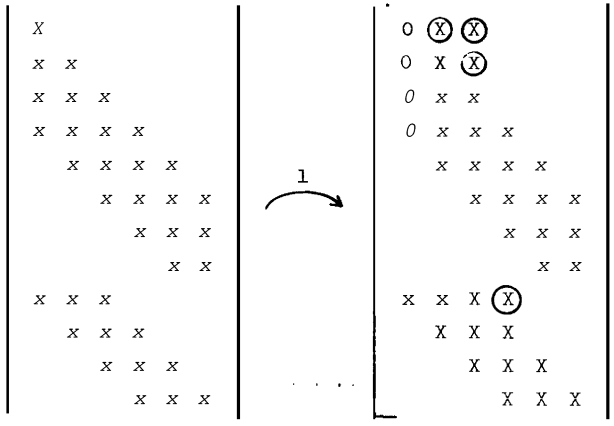
$$\begin{pmatrix} A \\ C \end{pmatrix} x \approx \begin{pmatrix} b \\ d \end{pmatrix} \tag{5.8}$$

where now A and C are band matrices. We assume that C has a band width which is smaller or equal to the band width of A . Using Givens rotations we transform

$$G\begin{pmatrix} A \\ C \end{pmatrix} = \begin{pmatrix} 0' \\ A' \end{pmatrix} \tag{5.9}$$

where A' has the same bandwidth as A and $0'$ is zero up to some elements in the right bottom corner. Using the elements of 0' and A' we then compute x by backsubstitution.

In the k-th step of the algorithm we annihilate the k-th column of A and produce the k-th row of A' . We explain the rotation:: for the example where A has 4 and C has 3 diagonals.



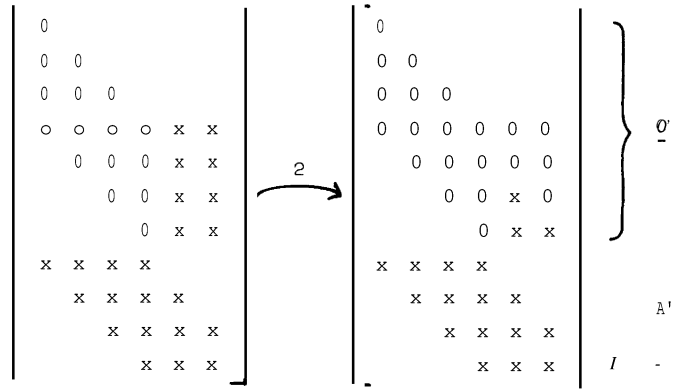
In step 1 we annihilate the elements of the first column of A using the first row of C . This produces the new non-zero elements (X) . In a second "cleaning" step we zero the elements (X) in the top part using the third row:

$$\begin{bmatrix} 0 & \textcircled{X} & \textcircled{X} \\ 0 & X & \textcircled{X} \\ 0 & X & X \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & \textcircled{X} & 0 \\ 0 & X & 0 \\ 0 & X & X \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & 0 & 0 \\ 0 & X & 0 \\ 0 & X & X \end{bmatrix}$$

After these two steps the remaining matrix is:

$$\begin{bmatrix} 0 & & & & \\ 0 & x & & & \\ 0 & x & x & & \\ 0 & x & x & x & \\ & x & x & x & x \\ & & X & X & X & \dot{X} \\ & & & x & x & x \\ & & & & x & x \\ x & x & x & x & & \\ & x & x & x & & \\ & & x & x & x & \\ & & & x & x & x \end{bmatrix}$$

Now we can use the second row of C to zero the second column of etc. A special treatment is needed at the border of the matrix. using the last row of C to zero out the n-5 column gives



In the cleaning step 2 we can zero all but 3 elements in the upper part of the matrix using the last row of A . Now the solution can be computed by backsubstitution: first x_{n-6} and x_{n-7} using Q' and the other unknown using A' . In Reinsch's "Smoothing with Spline Functions" [33] the problem

$$(Q^T D^2 Q + \lambda T)z = Q^T y \qquad (5.10)$$

subject to

$$\|D Q z\|^2 = S^2 \qquad (5.11)$$

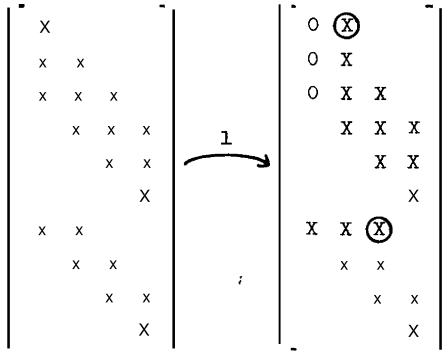
occurs, where $\lambda > 0$, D is an $(n+1) \times (n+1)$ diagonal matrix, Q an $(n+1) \times (n-1)$ tridiagonal matrix and T a positive definite tridiagonal $(n-1) \times (n-1)$ matrix. S is a given constant and y a given $(n+1)$ vector. Reinsch solves the normal equation (6.0) using the Cholesky decomposition of the penta-diagonal coefficient matrix. Using the above described algorithm we can solve (5.10) as a least squares problem without sacrificing the sparseness of the matrices. As preparation we compute the Cholesky decomposition of T :

$$T = \underline{B}^T \underline{B} \quad , \quad B \text{ upper bidiagonal}$$

and (5.10) then becomes

$$\begin{pmatrix} \underline{D} & \underline{Q} \\ \sqrt{\lambda}\,\underline{B} & \end{pmatrix} \underline{z} \approx \begin{pmatrix} \underline{D}^{-1}\underline{y} \\ \underline{0} \end{pmatrix} \quad . \tag{5.12}$$

The matrix of (5.12) has the form (n = 5)



Zeroing the first column in step 1 leaves one element (X) that has to be removed in the cleaning step.

5.3 Bidiagonalization.

If the matrices $\underline{A}$ and $\underline{C}$ are dense then for problems (P2), (P3) and (P1) (after performing Eldén's transformation, see Section 3.3) we have to solve a least square problem of the form

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix} \underline{x} \approx \begin{pmatrix} \underline{b} \\ \underline{0} \end{pmatrix} \tag{5.13}$$

for every new value of λ. If $\underline{P}$ and $\underline{Q}$ are orthogonal then an equivalent system to (5.13) is

$$\begin{pmatrix} \underline{P}^T & \underline{0} \\ \underline{0} & \underline{Q}^T \end{pmatrix} \begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix} \underline{Q}\,\underline{Q}^T \underline{x} \approx \begin{pmatrix} \underline{P}^T \underline{b} \\ \underline{0} \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \underline{P}^T \underline{A}\,\underline{Q} \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix} \underline{y} \approx \begin{pmatrix} \underline{c} \\ \underline{0} \end{pmatrix} \tag{5.14}$$

with $Y := \underline{Q}^T x$ and $c := \underline{P}^T \underline{b}$.

Now we can choose $\underline{P}$ and $\underline{Q}$ so that (5.14) is simpler to solve than (5.13). Moré [29] proposed to choose P and Q so that

$$\underline{B} := \underline{P}^T \underline{A}\,\underline{Q} = \begin{pmatrix} \underline{R} \\ 0 \quad 0 \end{pmatrix} \quad , \tag{5.15}$$

i.e., to perform the QR decomposition of A with column pivoting. However the solution of (5.14) in this case is still an n^3 process. It has been pointed out by More/ (private communication) that in his application [29] only $\ell \leq 2$ iterations are needed to solve the secular equation. Therefore it does not matter in this case if the solution of (5.14) is an n^3 process. In general however we will have to perform much more iterations (see Section 6). With about double the amount of work to perform the decomposition (5.15) we can bidiagonalize $\underline{A}$:

$$\underline{B} := \underline{P}^T \underline{A}\,\underline{Q} = \begin{pmatrix} \diagdown & 0 \\ 0 & \end{pmatrix} \tag{5.16}$$

as proposed by [8] and [12]. Then to solve (5.14) we can use the algorithm of Section 5.1 that requires only $\sim n$ multiplications.

Bidiagonalizing a matrix $\underline{A}$ is the first step in computing the singular value decomposition [48]. But the diagonalizing of B requires that $\underline{P}$ and $\underline{Q}$ be formed explicitly and updated after every QR-step which makes the process fairly expensive. To transform our problem to bidiagonal form we do not need P and Q explicitly. It is sufficient to store the vectors of the Householder matrices that perform the transformation.

Let $\underline{A}$ be an $m \times n$ matrix. Then we can bidiagonalize A in n steps [48]:

$$\underline{B} = \underbrace{\underline{P}_n \cdots \underline{P}_1}_{\underline{P}^T} \underline{A} \underbrace{\underline{Q}_1 \cdots \underline{Q}_{n-1}}_{\underline{Q}}$$

where

$$\underline{P}_j = \begin{pmatrix} \underline{I}_{j-1} & \underline{0} \\ \underline{0} & \tilde{\underline{P}}_j \end{pmatrix} \begin{matrix} \} j-1 \\ \} m-j+1 \end{matrix}$$

$$\tilde{\underline{P}}_j = \underline{I} - \underline{w}_j \underline{w}_j^T \qquad \text{with} \quad \|\underline{w}_j\| = \sqrt{2}$$

and

$$\underline{Q}_j = \begin{pmatrix} \underline{I}_j & \underline{0} \\ \underline{0} & \tilde{\underline{Q}}_j \end{pmatrix} \begin{matrix} \} j \\ \} n-j \end{matrix}$$

$$\tilde{\underline{Q}}_j = \underline{I} - \underline{v}_j \underline{v}_j^T \qquad \text{with} \quad \|\underline{v}_j\| = \sqrt{2} \quad .$$

The transformation vectors $\underline{w}_j$ and $\underline{v}_j$ are stored in the matrix A as follows.

$$A = \begin{array}{|c|c|c|} \hline & \underline{v}_1^T & \\ \hline & & \underline{v}_2^T \\ \hline \underline{w}_1 & \underline{w}_2 & \ddots \\ \hline \end{array} \qquad (5.17)$$

To compute the vectors

$$\underline{P}\,\underline{x} \quad , \quad \underline{P}^T \underline{x} \quad , \quad \underline{Q}\,\underline{y} \quad , \quad \underline{Q}^T \underline{y}$$

we do not need $\underline{P}$ and $\underline{Q}$ explicitly. In (5.17) we have all the information to compute the operator $\underline{P}$, $\underline{P}^T$, $\underline{Q}$ or $\underline{Q}^T$ that act on a certain vector.

The computation of $\underline{y} = \underline{P}^T \underline{x}$, i.e., is done by the following procedure:

```
procedure ptx(m,n,a,x,y);
value m,n; integer n,m;
arrayx,   y;
begin integer i,k; real s;
     for i := 1 step 1 until m do y[i] := x[i];
 t: for i := 1 step 1 until n do
     begin
          s := 0;
          for k := i step 1 until m do s := s + a[k,i] x y[k];
          for k := i step 1 until m do y[k] := y[k] - a[k,i] x s;
     end i;
end
```

If we wish to compute $\underline{y} = P\underline{x}$, the only change we have to make in the above procedure is

```
t : for i := n step -1 until 1 do
```

The following procedure bidia bidiagonalizes $\underline{A}$, i.e., computes the diagonal $\underline{q}$ and the superdiagonal e of B and stores the transformation vectors $\underline{w}_{.j}$ and $\underline{v}_{.j}$ in $\underline{A}$ (5.17):

```
procedure bidia(m,n,a,q,e);
integer m,n; ______ m,n;
array a,q,e;
begin
    integer i,j,k; real s, fak;
    for i := 1 step 1 until n do
    begin
        comment transform (a_ii,...,a_mi) to (q_i,0,...,0);
        s := 0;
        for j := i step 1 until m do s := s + a[j,i]↑2;
        if s = 0 then q[i] := 0 else
        begin
            s := sqrt(s);
            q[i] := if a[i,i] > 0 then -s else s;
            fak := sqrt(s × (s + abs(a[i,i])));
            a[i,i] := a[i,i] - q[i];
            for k := i step 1 until m do a[k,i] := a[k,i]/fak;
            for j := i+1 step 1 until n do
```



```
            begin
                s := 0;
                for k := i step 1 until m do s := s + a[k,i] × a[k,j];
                fork := i step 1 until m do a[k,j] := a[k,j] - a[k,i] × s;
            end j;
        end s;
        comment transform (a_i,i+1,...,a_in) to (e_i,0, . . .,0);
        if i = n then goto ende;
        s := 0;
        for j := i+1 step 1 until n do s := s + a[i,j]↑2;
        if s = 0 then e[i] := 0 else
        begin
            s := sqrt(s);
            e[i] := if a[i,i+1] > 0 then -s else s;
            fak := sqrt(s × (s + abs(a[i,i+1])));
            a[i,i+1] := a[i,i+1] - e[i];
            for k := i+1 step 1 until n do a[i,k] := a[i,k]/fak;
            for j := i+1 step 1 until m do
            begin
                s := 0;
                for k := i+1 step 1 until n do s := s + a[j,k] × a[i,k];
                for k := i+1 step 1 until n do a[j,k] := a[j,k] - a[i,k] × s;
            end j;
        end s;
    end i;
ende:
end bidia;
```

The decomposition using bidia requires $\sim 2(m\,n^2 - n^3/3)$ multiplications. If $m > n$ it is possible to bidiagonalize A even cheaper [26], [5 I, [8]. The idea is to transform A first to an upper triangular matrix R and then bidiagonalize R. In this case the operation count is $\sim m\,n^2 + n^3$. An alternative approach to bidiagonalizing A has been suggested by Golub and Kahan [14]. This algorithm uses the matrix A not explicitly. Only the operator $A\,x$ is needed. Unfortunately it is not numerically stable but nevertheless it seems to be useful for sparse matrices [31].

5.4 Computation of the Derivatives of the Length Function.

If we want to solve the secular equation

$$f(\lambda) = \alpha^2 \tag{5.18}$$

using some high order iteration method we have to compute the derivatives of f. For the general problem (P1) we have

$$(A^TA + \lambda C^TC)x(\lambda) = A^Tb + \lambda C^Td \tag{5.19}$$

$$f(\lambda) = \|Cx(\lambda) - d\|^2 \quad .$$

By differentiating (5.19) we get

$$(A^TA + \lambda C^TC)x' = -C^T(Cx - d) \tag{5.20}$$

$$(A^TA + \lambda C^TC)x'' = -2\,C^TCx' \quad .$$

In general we have for $k \geq 2$

$$(A^TA + \lambda C^TC)x^{(k)} = -k\,C^TC\,x^{(k-1)} \quad . \tag{5.21}$$

Lemma 5.1. If $B_\lambda := A^TA + \lambda C^TC$ and $x^{(k)}$ is the solution of (5.21) then

(i) $$x^{(k)T} B_\lambda x^{(k)} = \begin{cases} x^T(A^Tb + \lambda C^Td), & k = 0 \\ -x'^T C^T(Cx - d), & k = 1 \\ -k\,x^{(k)T} C^TC\,x^{(k-1)}, & k \geq 2 \end{cases}$$

(ii) $$x^{(k)T} B_\lambda x^{(k)} = \begin{cases} \frac{1}{2} x^T B_\lambda x'' + x'^T C^Td, & k = 1 \\ \frac{k}{k+1} x^{(k-1)} B_\lambda x^{(k+1)}, & k \geq 2 \end{cases}$$

(iii) $$\|Cx^{(k)}\|^2 = \begin{cases} -x^T B_\lambda x' + x^T C^Td, & k = 0 \\ -\frac{1}{k+1} x^{(k)T} B_\lambda x^{(k+1)}, & k > 1 \end{cases}$$

(iv) $$\|Cx^{(k)}\| = \begin{cases} \frac{1}{2} x^{(2)T} C^T(Cx - d), & k = 1 \\ \frac{k}{k+1} x^{(k+1)T} C^TC\,x^{(k-1)}, & k > 2 \end{cases}$$

(v) $$\frac{d}{d\lambda}(x^{(k)T} B_\lambda x^{(k)}) = \begin{cases} -\|Cx - d\|^2 + \|d\|^2, & k = 0 \\ -(2k+1)\|Cx^{(k)}\|^2, & k \geq 1 \end{cases}$$

(vi) $$\frac{d}{d\lambda}\|Cx^{(k)}\|^2 = \begin{cases} -2\,x'^T B_\lambda x' + 2\,x'^T C^Td, & k = 0 \\ -\frac{2}{k+1} x^{(k+1)T} B_\lambda x^{(k+1)}, & k > 1 \end{cases}$$

Proof. Equations (i) follow, by multiplying (5.19), (5.20), (5.21) by $x^{(k)T}$ from the left.

Looking at two consecutive equations of (5.21):

$$B_\lambda x^{(k)} = - k C^T C x^{(k-1)} \tag{5.22}$$

$$B_\lambda x^{(k+1)} = - (k+1) C^T C x^{(k)} . \tag{5.23}$$

We obtain (ii) by multiplying the first (5.22) by $x^{(k)T}$ and the second by $x^{(k-1)T}$ and by eliminating the expression $x^{(k)T} C^T C x^{(k-1)}$. To prove (iii) we multiply the equation for $x^{(k+1)}$ of (5.21) by $x^{(k)}$ and (5.20) by x^T. Equation (iv) follows by multiplying (5.22) by $x^{(k+1)T}$ and (5.23) by $x^{(k)}$ and subtracting both equations.

Finally observe that $B_\lambda' = C^T C$, therefore

$$\frac{d}{d\lambda} (x^{(k)T} B_\lambda x^{(k)})$$

$$= x^{(k+1)T} B_\lambda x^{(k)} + x^{(k)T} \{B_\lambda' x^{(k)} + B_\lambda x^{(k+1)}\}$$

$$= 2 x^{(k+1)T} B_\lambda x^{(k)} + \|Cx^{(k)}\|^2 .$$

Now for k = 0 using (iii) we have

$$\frac{d}{d\lambda} (x^T B_\lambda x) = - \|Cx\|^2 + 2 x^T C^T d = - \|Cx - d\|^2 + \|d\|^2 .$$

For $k \geq 1$ using (iii) we get

$$\frac{d}{d\lambda} (x^{(k)T} B_\lambda x^{(k)}) = - (2k+1) \|Cx^{(k)}\|^2 .$$

Using (i) we can write

$$\frac{d}{d\lambda} \|Cx^{(k)}\|^2 = - \frac{2}{k+1} x^{(k+1)T} B_\lambda x^{(k+1)} , \qquad \square$$

Corollary 5.1. If x is the solution of (5.19) then

$$\left.\begin{aligned} \int f(\lambda) d\lambda &= \lambda \|d\|^2 - x^T (A^T A + \lambda C^T C) x + \text{const.} \\ &= - A d^T (Cx - d) - b^T Ax + \text{const.} \end{aligned}\right\} \tag{5.24}$$

Proof. From equation (v) of Lemma 5.1 we have

$$f(\lambda) = \|Cx - d\|^2 = \|d\|^2 - \frac{d}{d\lambda} (x^T B_\lambda x) .$$

By integrating and using (5.19) the result follows immediately. $\square$

Corollary 5.2. If $B_\lambda = A^T A + \lambda C^T C$ then for $k \geq 1$,

(i) $$\frac{d}{d\lambda} (x^{(k)T} B_\lambda x^{(k)}) = \frac{2k+1}{k+1} x^{(k)T} B_\lambda x^{(k+1)}$$

(ii) $$\frac{d}{d\lambda} (x^{(k)T} B_\lambda x^{(k+1)}) = 2 x^{(k+1)T} B_\lambda x^{(k+1)}$$

Corollary 5.3. If $x^{(k)}$ is a solution of (5.21), then for $k \geq 1$

(i) $$\frac{d}{d\lambda} \|Cx^{(k)}\|^2 = 2 x^{(k+1)T} C^T C x^{(k)}$$

(ii) $$\frac{d}{d\lambda} (x^{(k+1)T} C^T C x^{(k)}) = \frac{2k+3}{k+1} \|Cx^{(k+1)}\|^2 .$$

Theorem 5.1. Let x, x', and $x^{(k)}$ be solutions of (5.19), (5.20), and (5.21) and let $B_\lambda = A^T A + \lambda C^T C$. Then

$$f(A) = \|Cx - d\|^2 = - x^T B_\lambda x' - d^T (Cx - d) \tag{5.25}$$

$$f'(A) = 2 x'^T C^T (Cx - d) = - 2 x'^T B_\lambda x' \tag{5.26}$$

and for $k \geq 1$

$$f^{(2k)}(\lambda) = (k+1)\gamma_{2k} \|\underline{C}\underline{x}^{(k)}\|^2 = -\gamma_{2k}\, \underline{x}^{(k)}\, \underline{B}_\lambda\, \underline{x}^{(k+1)} \tag{5.27}$$

$$f^{(2k+1)}(\lambda) = (k+1)\gamma_{2k+1}\, \underline{x}^{(k)T}\, \underline{C}^T\underline{C}\, \underline{x}^{(k+1)}$$

$$= -\gamma_{2k+1}\, \underline{x}^{(k+1)T}\, \underline{B}_\lambda\, \underline{x}^{(k+1)} \tag{5.28}$$

where

$$\left.\begin{aligned} \gamma_{2k} &= \frac{1\cdot 3\cdot 5\cdot \ldots \cdot (2k+1)}{(k+1)!}\, 2^k \\ \gamma_{2k+1} &= \frac{1\cdot 3\cdot 5\cdot \ldots \cdot (2k+1)}{(k+1)!}\, 2^{k+1} \end{aligned}\right\} \tag{5.29}$$

Proof. The proof is by induction. If

$$f(\lambda) = \|\underline{C}\underline{x} - \underline{d}\|^2$$

then by differentiating and by Lemma 5.1 (i) we have

$$f'(A) = 2\, \underline{x}'^T\, \underline{C}^T(\underline{C}\underline{x} - \underline{d}) = -2\, \underline{x}'^T\, \underline{B}_\lambda\, \underline{x}' .$$

Differentiating again using (v) and (iii) of Lemma 5.1 we get

$$f''(A) = 2\cdot 3\, \|\underline{C}\underline{x}'\|^2 = -3\, \underline{x}'^T\, \underline{B}_\lambda\, \underline{x}''$$

which is (5.27) for $k = 1$. Now we can use Corollary 5.2 and 5.3 to compute the higher derivatives.

Assume

$$f^{(2k)}(\lambda) = (k+1)\gamma_{2k}\, \|\underline{C}\underline{x}^{(k)}\|^2 = -\gamma_{2k}\, \underline{x}^{(k)}\, \underline{B}_\lambda\, \underline{x}^{(k+1)} .$$

Then

$$f^{(2k+1)}(\lambda) = (k+1)\gamma_{2k}\cdot 2\, \underline{x}^{(k)T}\, \underline{C}^T\underline{C}\, \underline{x}^{(k+1)}$$

$$= -\gamma_{2k}\cdot 2\, \underline{x}^{(k+1)T}\, \underline{B}_\lambda\, \underline{x}^{(k+1)}$$

which is (5.28) for $\gamma_{2k+1} = 2\gamma_{2k}$. Again differentiating using Corollary 5.2 and 5.3 yields

$$f^{(2k+2)}(\lambda) = \gamma_{2k+1}(2k+3)\, \|\underline{C}\underline{x}^{(k+1)}\|^2$$

$$= -\gamma_{2k+1}\frac{2k+3}{k+2}\, \underline{x}^{(k+1)T}\, \underline{B}_\lambda\, \underline{x}^{(k+2)}$$

which is (5.27) for $k+1$ and $\lambda_{2k+2} = \gamma_{2k+1}\frac{2k+3}{k+2}$. From this recursion for γ_i it is easy to verify (5.29). □

Theorem 5.1 shows that we can compute cheaply derivatives of f. To compute $x^{(k)}$ we have to solve a linear system with the same matrix $\underline{B}_\lambda$ as for $\underline{x}$. Therefore we can use a factorization of $\underline{B}_\lambda$, If A. > 0 then $\underline{x}, \underline{x}', \ldots, \underline{x}^{(k)}$ is a solution of the least squares problem

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix}\underline{x} \approx \begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\,\underline{d} \end{pmatrix}$$

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix}\underline{x}' \approx -\frac{1}{\sqrt{\lambda}}\begin{pmatrix} \underline{0} \\ \underline{C}\underline{x} - \underline{d} \end{pmatrix}$$

$$\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix}\underline{x}^{(k)} \approx -\frac{k}{\sqrt{\lambda}}\begin{pmatrix} \underline{0} \\ \underline{C}\underline{x}^{(k-1)} \end{pmatrix} \quad , \quad k > 2 .$$

If we transform by an orthogonal matrix $\underline{G}$

$$\underline{G}\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix} = \begin{pmatrix} \underline{R}_\lambda \\ \underline{0} \end{pmatrix} \quad , \quad \underline{R}_\lambda \text{ upper triangular,}$$

then $\underline{R}_\lambda^T \underline{R}_\lambda$ is the **Cholesky decomposition** of $\underline{B}_\lambda = \underline{A}^T\underline{A} + A\,\underline{C}^T\underline{C}$. To solve

$$\underline{B}_\lambda\, \underline{x}^{(k)} = - k\, \underline{C}^T\underline{C}\, \underline{x}^{(k-1)}$$

we have two possibilities

(1) **compute** $\underline{y}^{(k)}$ by forward substitution **from**

$$\underline{R}_\lambda^T\, \underline{y}^{(k)} = - k\, \underline{C}^T(\underline{C}\underline{x}^{(k-1)}) \;;$$

(2) compute $\underline{y}^{(k)}$ using $\underline{G}$

$$\begin{pmatrix} \underline{y}^{(k)} \\ \underline{k} \end{pmatrix} = \underline{G}\begin{pmatrix} \underline{0} \\ -\frac{k}{\sqrt{\lambda}}\,\underline{C}\underline{x}^{(k-1)} \end{pmatrix} .$$

Then we obtain $x^{(k)}$ by back substitution in

$$\underline{R}_\lambda\, \underline{x}^{(k)} = \underline{y}^{(k)} .$$

If we define $\underline{z}^{(k)} := \underline{C}\underline{x}^{(k)}$ then

$$f^{(2k-1)}(A) = k\, \gamma_{2k-1}\, \underline{z}^{(k-1)T}\, \underline{z}^{(k)}$$

or equivalently

$$f^{(2k-1)}(\lambda) = - \gamma_{2k-1}\, \|\underline{y}^k\|^2 .$$

Similarly

$$f^{(2k)}(A) = (k+1)\gamma_{2k}\, \|\underline{z}^{(k)}\|^2$$

or if $\underline{y}^{(k+1)}$ has been computed

$$f^{(2k)}(\lambda) = - \gamma_{2k}\, \underline{y}^{(k)T}\, \underline{y}^{(k+1)} .$$

In Section 6 we **shall** consider third order iteration methods to solve the secular equation for $\lambda > 0$. We need therefore the values of **f ,** f' , and f" , which can be computed as follows:

1. **Compute** $\underline{G}$ and $\underline{R}_\lambda$ so that

$$\underline{G}\begin{pmatrix} \underline{A} \\ \sqrt{\lambda}\,\underline{C} \end{pmatrix} = \begin{pmatrix} \underline{R}_\lambda \\ \underline{0} \end{pmatrix} .$$

2. **Compute** $\underline{y}$ from

$$\underline{R}_\lambda^T\, \underline{y} = \underline{A}^T\underline{b} + \lambda\, \underline{C}^T\underline{d}$$

or using $\underline{G}$ by

$$\begin{pmatrix} \underline{y} \\ \underline{h} \end{pmatrix} = \underline{G}\begin{pmatrix} \underline{b} \\ \sqrt{\lambda}\,\underline{d} \end{pmatrix} .$$

3. **Compute** x **from** $\underline{R}_\lambda\, \underline{x} = \underline{y}$ **and form** $\underline{z} := \underline{C}\underline{x} - \underline{d}$.

4. $f(\lambda) = \|\underline{z}\|^2$.

5. Compute $\underline{y}'$ by solving

$$\underline{R}_\lambda^T\, \underline{y}' = - \underline{C}^T\underline{z}$$

or by using G

. . .

$$\begin{pmatrix} \underline{y}' \\ \underline{h}' \end{pmatrix} := \underline{G}\begin{pmatrix} \underline{0} \\ -\frac{1}{\sqrt{\lambda}}\,\underline{z} \end{pmatrix}$$

6. $f'(A) := -2\,\|\underline{y}'\|^2$.

7. Compute x' from $\underline{R}_\lambda\,\underline{x}' = \underline{y}'$ and form $z' := C\underline{x}'$.

8. $f''(A) = 6\,\|\underline{z}'\|^2$.

If we do not want to store $\underline{G}$ then we have to compute $\underline{y}'$ in step 5 by forward substitution. However, y in step 2 can be computed together with the decomposition in step 1 without forming $\underline{G}$ explicitly. Eldén gives in [8] similar recursions to compute the derivatives.

6. One-point Iteration Methods to Solve the Secular Equation.

In this chapter we discuss how to find the solution $A^* > 0$ of the secular equation

$$f(\lambda) = \alpha^2 \tag{6.1}$$

that we need to solve problem (P1), (P2) or (P3). The length function f is a positive rational function for all A and decreasing in $(0,\infty)$. If we start with A = 0 Newton's method will produce a strictly increasing sequence $\{\lambda_n\}$ which converges globally to A^* . However as it was observed in [34] if α is small, convergence is slow and as a remedy Reinsch suggested solving the equation

$$\frac{1}{\sqrt{f}} - \frac{1}{\alpha} = 0 \tag{6.2}$$

instead of $\sqrt{f} - \alpha = 0$ using Newton's method. Indeed convergence is much better in this case. There are several possible interpretations and explanations for this fact. For our purpose we compare the Newton step resulting for the three equations

$$g_1(\lambda) := f(A) - \alpha^2 = 0$$

$$g_2(\lambda) := \sqrt{f(\lambda)} - \alpha = 0$$

$$g_3(\lambda) := \frac{1}{\sqrt{f(\lambda)}} - \frac{1}{\alpha} = 0 \quad .$$

A short calculation yields the following Newton iteration functions for the three equations:

$$\lambda - \frac{f - \alpha^2}{f'} \qquad \text{for } g_1 , \tag{6.3}$$

$$\lambda - \frac{f - \alpha^2}{f'} \; \frac{2}{1 + \frac{\alpha}{\sqrt{f}}} \qquad \text{for } g_2 , \tag{6.4}$$

$$\lambda - \frac{f - \alpha^2}{f'} \; \frac{2\,\frac{\sqrt{f}}{\alpha}}{1 + \frac{\alpha}{\sqrt{f}}} \qquad \text{for } g_3 . \tag{6.5}$$

For $\lambda = 0$ we typically have $f(0) \gg \alpha^2$. Since $f' < 0$ for $A > 0$ the step in (6.4) will be about twice as big as in (6.3). But (6.5) will produce an even larger step, proportional to $\frac{\sqrt{f}}{\alpha}$ the discrepancy of $\sqrt{f}$ and α. If $f \approx \alpha^2$ then all the three steps are of about the same size.

We can look at the two functions

$$h_1(\lambda) = 2 \Big/ \left(1 + \frac{\alpha}{\sqrt{f(\lambda)}}\right)$$

$$h_2(\lambda) = h_1(\lambda)\,\frac{\sqrt{f}}{\alpha}$$

as functions that help to accelerate the global convergence of (6.3) by preserving the order (all iterations are Newton sequences and of second order).

6.1 Convergence Factors.

For the following discussion we change notation. Let f be a given function. We are looking for a number s such that $f(s) = 0$. We assume

that f has sufficient continuous derivatives in a neighborhood of s and furthermore we assume that s is a simple zero of f. We consider one point iteration methods without memory [43]

$$\left\{ \begin{array}{l} x_0 \text{ arbitrary} \\ x_{n+1} = F(x_n) \qquad n = 0,1,\ldots \end{array} \right\} \tag{6.6}$$

where

$$F(x) = x - \frac{f(x)}{f'(x)}\,G(x) \tag{6.7}$$

and $G(x)$ is an appropriate chosen function which we will call the "convergence factor". The idea is to choose G so that the global convergence using (6.7) is better than Newton's iteration $(G(x) \equiv 1)$. As has been pointed out by Kahan [24], every sequence generated by an iteration (6.6) can be interpreted as a sequence obtained by applying Newton's method to a certain equation

$$g(x) = 0 . \tag{6.8}$$

Indeed if we put

$$x - \frac{g(x)}{g'(x)} = F(x)$$

a short calculation yields (with some $c \neq 0$)

$$g(x) = c \cdot \exp\left(\int \frac{dx}{x - F(x)}\right) . \tag{6.9}$$

Solving (6.8) with (6.9) using Newton's method yields the sequence (6.6). Some examples may illustrate the point.

(1) Let $x_{n+1} = 1 + \frac{x_n}{2}$, $x_0 = 0$. This sequence converges linearly to $s = 2$. Indeed using (6.9) we get

$$g(x) = \exp\left(\int dx \Big/ \left(\frac{x}{2} - 1\right)\right) = \left(\frac{x}{2} - 1\right)^2 .$$

$g(x)$ has a double zero $s = 2$ and therefore Newton's iteration converges linearly.

(2) Consider the Halley iteration formula $x_{n+1} = F(x_n)$ with

$$F(x) = x - \frac{2 f(x) f'(x)}{2(f'(x))^2 - f''(x) f(x)} .$$

using (6.9) we obtain

$$g(x) = \exp\left(\int \left(\frac{f(x)}{f'(x)} - \frac{f''(x)}{2f'(x)}\right) dx\right) = \frac{f(x)}{\sqrt{f'(x)}} .$$

Therefore using Halley's method for $f(x) = 0$ is applying Newton's method to

$$g(x) = \frac{f(x)}{\sqrt{f'(x)}} = 0 \quad , [3].$$

(3) Consider the iteration

$$x_{n+1} = x_n - \frac{f(x_n) - \alpha}{f'(x_n)} \cdot \frac{f(x_n)}{\alpha} \tag{6.10}$$

to solve $f(x) = \alpha$. Using (6.9) we get

$$g(x) = 1 - \frac{\alpha}{f(x)}$$

or since g is only determined to a constant we may also divide by $-a$ and get

$$g(x) = \frac{1}{f(x)} - \frac{1}{\alpha} .$$

Therefore (6.10) is obtained by applying Newton's method to

$$\frac{1}{f(x)} - \frac{1}{\alpha} = 0$$

instead of $f(x) - \alpha = 0$.

For the class of iteration functions $F(x)$ (6.7) we would like to consider however if it will not be possible in general to evaluate the integral for g in (6.9) explicitly and thus provide a nice explanation of the iteration function.

The order of convergence [22] of an iteration formula

$$x_{n+1} = F(x_n)$$

is

one (or linear convergence), if $|F'(s)| < 1$

two (or quadratic convergence), if $F'(s) = 0$

three (or cubic convergence), if $F'(s) = F''(s) = 0$

m , if $F'(s) = F''(s) = \ldots = F^{(m-1)}(s) = 0$.

For $G(x) \equiv 1$ the iteration with

$$F(x) = x - \frac{f(x)}{f'(x)} \cdot G(x) \tag{6.11}$$

is Newton's method and it is of second order for a zero of f with multiplicity one. Differentiation (6.11) gives

$$F'(x) = 1 - \left(\frac{f(x)}{f'(x)}\right)' G(x) - \frac{f(x)}{f'(x)} G'(x) .$$

Since $f(s) = 0$ we have $F'(s) = 0$ if $G(s) = 1$ and $f(s)\cdot G'(s) = 0$.

Therefore we have

Lemma 6.1. Let G be differentiable, $G(s) = 1$, $f(s)\cdot G'(s) = 0$.

Then the iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\, G(x_n)$$

is of second order for simple zeros s of f . Since s is unknown we have to choose

$$G(x) = H(f(x), f'(x), \ldots) \ .$$

Examples:

(4) $\quad G(x) = H(f) := 1 + f(x)$

$$\Rightarrow \quad F(x) = x - \frac{f(x)}{f'(x)}\,(1 + f(x)) \ . \tag{6.13}$$

(6.13) yields a second order iteration formula Indeed using (6.9) we see that the iteration is obtained solving $g(x) = 0$ with Newton where

$$g(x) = \exp\left(\int \frac{f'(x)dx}{f(x)(1 + f(x))} \right) = \frac{1}{1 + f(x)} - 1 \ .$$

(6.13) is a special case $(a = 1)$ of

(5) $\quad H(f) = \dfrac{\alpha + f}{\alpha}$ (6.14)

for some $\alpha \neq 0$. Therefore the iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\ \frac{\alpha + f(x_n)}{\alpha}$$

is the same as applying Newton's method to

$$g(x) = \frac{1}{f(x) + \alpha} - \frac{1}{\alpha} = 0 \ , \quad \alpha \neq 0 \ .$$

(6) $\quad G(x) = 1 \Big/ \left(1 - \frac{1}{2}\, \frac{f(x)f''(x)}{(f'(x))^2} \right)$.

This choice is Halley's method. Again we have $G(s) = 1$.

6.2 Third Order Iterative Methods.

We consider again the iteration

$$x_{n+1} = F(x_n) = x_n - \frac{f(x_n)}{f'(x_n)} \cdot G(x_n) \ . \tag{6.15}$$

In Section 6.1 we saw that if $G(s) = 1$ then (6.15) is quadratic convergent. Let $u(x) := \dfrac{f(x)}{f'(x)}$. Then

$$F(x) = x - u(x) \cdot \ldots \tag{6.16}$$

and we can ask for the conditions that the iteration (6.15) be cubically convergent. Differentiating (6.16) we get (dropping the argument)

$$F' = 1 - u'G - uG'$$

$$F'' = - u''G - 2u' G' - uG''$$

$$u = f/f'$$

$$u' = 1 - \frac{ff''}{f'^2}$$

$$u'' = - \frac{f''}{f'} + 2\,\frac{ff''}{f'^3} - \frac{ff'''}{f'^2} \ .$$

Now we want to have

$$F'(s) = F''(s) = 0 \ . \tag{6.17}$$

Since $u(s) = 0$, $u'(s) = 1$ and $u''(s) = -\frac{f''(s)}{f'(s)}$, this is the case if

$$G(s) = 1 \quad \text{and} \quad G'(s) = \frac{1}{2}\frac{f''(s)}{f'(s)} \ . \tag{6.18}$$

Theorem 6.1. Let $H \in C^2[-a,a]$ for some $a > 0$. The iteration $x_{n+1} = F(x_n)$ with

$$F(x) = x - \frac{f(x)}{f'(x)} \cdot H\left(\frac{f(x)f''(x)}{(f'(x))^2}\right) \tag{6.19}$$

converges cubically to a simple zero of f if and only if

$$H(0) = 1 \quad \text{and} \quad H'(0) = \frac{1}{2} \ .$$

Proof. The function F in (6.19) is the special case of (6.16) with

$$G(x) = H\left(\frac{f(x)f''(x)}{(f'(x))^2}\right) \ .$$

Let $t(x) = \frac{f(x)f''(x)}{(f'(x))^2}$. Then $t(x) = 1-u'(x)$ and therefore

$$t'(s) = -u''(s) = \frac{f''(s)^3}{f'(s)^3} \ .$$

Clearly $G(s) = 1 \Leftrightarrow H(0) = 1$ and

$$G'(s) = H'(t(s))t'(s)$$

$$= H'(0) \cdot \frac{f''(s)}{f'(s)} = \frac{1}{2}\frac{f''(s)}{f'(s)}$$

$$\Leftrightarrow H'(0) = \frac{1}{2} \ . \quad \square$$

Many well known third order iterative methods are special cases of Theorem 6.1. Let

$$t := \frac{f(x)f''(x)}{(f'(x))^2} \tag{6.20}$$

(1) Euler's formula.

$$H(t) = \frac{2}{1+\sqrt{1-2t}} = 1 + \frac{1}{2}t + \frac{1}{2}t^2 + \frac{5}{8}t^3 + \ldots$$

clearly $H(0) = 1$ and $H'(0) = 1/2$.

(2) Halley's formula.

$$H(t) = \frac{1}{1-\frac{1}{2}t} = 1 + \frac{1}{2}t + \frac{1}{4}t^2 + \frac{1}{8}t^3 + \cdots \ .$$

(3) Quadratic inverse interpolation [43].

$$H(t) = 1 + \frac{1}{2}t \ .$$

(4) Ostrowski's square root iteration [30].

$$H(t) = \frac{1}{\sqrt{1-t}} = 1 + \frac{1}{2}t + \frac{3}{8}t^2 + \frac{5}{16}t^3 + \cdots \ .$$

(5) Hansen-Patrick family [19].

$$H(t) = \frac{\alpha+1}{\alpha+\sqrt{1-(\alpha+1)t}} = 1 + \frac{1}{2}t + \frac{\alpha+3}{8}t^2 + \cdots \ .$$

(6) To solve the secular equation we will use

$$H(t) = e^{\frac{1}{2}t} = 1 + \frac{1}{2}t + \frac{1}{8}t^2 + \frac{1}{48}t^3 + \ldots$$

This formula has the advantage that $H(t) > 0$ even if the starting point is far away from the solution s . For large t , the other formulas may

not work (wrong sign in H , argument of the square root is negative).

The following lemma is useful for construction of third order iteration methods.

Lemma 6.2. Let $H_1(t)$ and $H_2(t)$ be two functions with

$$H_i(0) = a \quad \text{and} \quad H_i'(0) = b \ , \quad i = 1,2 \ .$$

Then the three mean functions

$$A = (H_1 + H_2)/2$$

$$B = \sqrt{H_1 H_2}$$

$$C = 2\Big/\left(\frac{1}{H_1} + \frac{1}{H_2}\right)$$

have the same property.

Proof. Obviously $A(0) = B(0) = C(0) = a$.

$$A' = (H_1' + H_2')/2 = b \ .$$

$$B' = (H_1'H_2 + H_1H_2') \,/\, (2\sqrt{H_1H_2})$$

$$B'(0) = (b \cdot a + a \cdot b) / (2a) = b \ .$$

$$C = \frac{2 H_1 H_2}{H_1 + H_2}$$

$$C' = 2\,\frac{(H_1 + H_2)(H_1'H_2 + H_1H_2') - (H_1' + H_2')H_1H_2}{(H_1 + H_2)^2}$$

$$C'(0) = 2\,\frac{2a(2ab) - 2ba^2}{4a^2} = b \ . \quad \square$$

Any of the three means of the examples (1) to (6) yields a new third order iterative method.

Notice that not every third order iterative method must have the form (6.19). For example, consider

$$G(x) = H(t(x)) + f^2(x) \ ,$$

with $H(0) = 1$, $H'(0) = 1/2$, and $t(x)$ defined by (6.20). Clearly this choice of the convergence factor G leads to a third order formula which has not the form of Theorem 6.1.

However it is possible to describe al the third order iteration methods. Let

$$F_k := x + \sum_{i=1}^{k-1} (-1)^i \frac{f^i}{i!} \frac{1}{f'} \left(\frac{d}{dx}\right)^{i-1} \frac{1}{f'}$$

denote the Schrbder iteration functions (see [7] or [21]). The iteration

$$x_{n+1} = F_k(x_k)$$

is of k-th order. Now every k-th order iteration function can be written as

$$F(x) = F_k(x) + f^k(x).\varphi(x)$$

where φ is an arbitrary function [7],[43]. Therefore for $k = 2$ we obtain the most general G for a third order method

$$G(x) = 1 + \frac{1}{2}\, t(x) \cdot f^2(x) \cdot \varphi(x)$$

which we can write

$$G(x) = H(t(x)) + f^2(x) \cdot \psi(x)$$

with arbitrary ψ .

6.3 The Convergence Factor for a Third Order Method.

Assume that $x_{n+1} = x_n - K(x_n)$ is a third order iterative method for solving $f(x) = 0$. We clearly have

$$K(s) = K''(s) = 0 \ , \quad K'(s) = 1 \ . \tag{6.21}$$

Now consider the iteration $x_{n+1} = F(x_n)$ with

$$F(x) = x - K(x) \cdot G(x) \ . \tag{6.22}$$

We have

$$\left.\begin{aligned} F' &= 1 - K'G - KG' \\ F'' &= \ - K''G - 2K'G' - KG'' \ . \end{aligned}\right\} \tag{6.23}$$

Lemma 6.3. The iteration $x_{n+1} = F(x_n)$ with F defined in (6.22) is also of third order if $G \in C^2[a,b]$ with $s \in (a,b)$ and

$$G(s) = 1 \ , \quad G'(s) = 0 \ . \tag{6.24}$$

Proof. If we use (6.21) and (6.24) in (6.23) then we have $F'(s) = F''(s) = 0$. □

If we choose G so that

$$G''(s) = -\frac{1}{3} K'''(s)$$

then we would have a fourth order iteration. However we think that a third order method with good global convergence is more useful than a local convergent fourth order formula.

The following functions are examples of possible convergence factors for a third order iteration:

(1) $$G(x) = (t^{\beta}(x) + t^{-\beta}(x))/2$$

where $t(x) = \dfrac{f(x) + \alpha}{\alpha} > 0$

and $\alpha, \beta \in R$.

(2) $$G(x) = \cosh(f(x) \cdot r(x))$$

where $r \in C^2[a,b]$ with $s \in (a,b)$.

(3) $$G(x) = L(f^2(x) \cdot r(x))$$

where $L \in C^2[-d,d]$ $d > 0$ and $L(0) = 1$,
$r \in C^2[a,b]$ with $s \in (a,b)$.

We know now how to choose convergence factors to preserve or increase the order of an iteration formula. Our aim is however to improve global convergence. Let

$$F(x) = x - K(x)G(x)$$

be the iteration function. Then G has to be chosen so that

$$g(x) = \exp\left(\int \frac{dx}{K(x) \cdot G(x)} \right) \tag{6.25}$$

is nearly linear if x is far away from s . Since the iteration $x_{n+1} = F(x_n)$ is the same as applying Newton's method to $g(x) = 0$, global convergence will be good if g is linear far away from the solution s . However, since (6.25) may be impossible to compute explicitly, this requirement seems not to be practical to determine G . We have to estimate G by other means. In Section 6.4 we shall interpret some G geometrically. Those functions can serve as models for others.

6.4 Geometrical Interpretation of the Convergence Factors.

It is possible to derive iterative methods as follows:

(i) choose a "simple" function h so that

$$f^{(i)}(x) = h^{(i)}(x) , \quad i = 0,1,\ldots,k ;$$

(ii) solve analytically $h(z) = \alpha$ obtaining $z = z(x)$;

(iii) use the iteration $x_{n+1} = z(x_n)$ to solve the equation $f(x) = a$.

The function h should be simple so that $h(z) = \alpha$ can be solved analytically. h approximates f and some derivatives locally at one point x and we thus obtain a one point iteration formula without memory.

Instead of approximating f one can also find iteration methods by approximating the inverse function locally.

(i) Choose a function h(y) so that

$$(f^{[-1]}(y))^{(i)} = h^{(i)}(y) , \quad i = 0,1,\ldots,k .$$

(ii) Put $y = f(x_n)$ and use the iteration formula

$$x_{n+1} = h(\alpha)$$

to solve $f(x) = \alpha$.

Since we do not know $f^{[-1]}$, the derivatives must be replaced using derivatives of f [21]:

$$(f^{[-1]}(y))' = \frac{1}{f'(f^{[-1]}(y))}$$

$$(f^{[-1]}(y))'' = -\frac{f''(f^{[-1]}(y))}{f'^3(f^{[-1]}(y))} .$$

If h approximates f resp $f^{[-1]}$ well we can expect a good global convergence. We give in the following some examples of methods derived by interpolation. The convergence factors of these examples may help to choose a method analytically.

(1) Newton's method $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ is obtained approximating f or $f^{[-1]}$ locally by a linear function h ($h(x) = ax+b$ resp. $h(y) = ay + b$) .

(2) We choose $h(z) = \frac{a}{(z-b)^2}$ and determine a, b so that $h^{(i)}(x) = f(i)(x)$, $i = 1,2$, giving (dropping the argument x):

$$\left.\begin{aligned} a &= 4\, f^3/f'^2 \\ b &= -x - 2\, f/f' . \end{aligned}\right\} \tag{6.26}$$

Now $h(z) = \alpha$ gives

$$z = -b \pm \sqrt{\frac{a}{\alpha}} .$$

using (6.26) we get

$$z = x - \frac{f}{f'} 2\left(\pm \sqrt{\frac{f}{\alpha}} - 1 \right) .$$

If we use only the + sign we have the iteration

$$\left.\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n) - \alpha}{f'(x_n)}\, G(x_n) \\ \text{with} \quad G(x) &= 2\sqrt{\frac{f}{\alpha}} \Big/ \left(1 + \sqrt{\frac{\alpha}{f}} \right) . \end{aligned}\right\} \tag{6.27}$$

But (6.27) is Reinsch's proposal to solve $\frac{1}{\sqrt{f}} - \frac{1}{\sqrt{\alpha}} = 0$ with Newton instead of $f-a = 0$. We know from Section 3 that the length function f

has the form

$$f(x) = \sum_{i=1}^{n} \frac{c_i^2}{(x - x_i)^2}$$

Therefore it is reasonable to approximate f by $h = a/(x-b)^2$.

(3) Instead of solving $f(x) - \alpha = 0$ with Newton's method we can also solve

$$g(x) = (f(x))^{-1/\beta} - \alpha^{-1/\beta} \tag{6.28}$$

if $f(x) > 0$ and $a > 0$. Newton's iteration for (6.28) yields the same iteration function as if we approximate f and f' locally by

$$h(x) = \frac{a}{(x+b)^{1/\beta}} .$$

The resulting iteration formula is

$$x_{n+1} = x_n - \frac{f(x_n) - \alpha}{f'(x_n)} \cdot G(x_n)$$

with

$$G(x) = \frac{\beta f}{f - \alpha}\left(\sqrt[\beta]{\frac{f}{\alpha}} - 1\right) . \tag{6.29}$$

If we let $\beta \to \infty$ in (6.29) we get

$$G(x) = \frac{f}{f-\alpha} \ln\left(\frac{f}{\alpha}\right) . \tag{6.30}$$

This convergence factor is also obtained by solving $\ln(f) - \ln(\alpha) = \ln(f/\alpha) = 0$ with Newton or by approximating locally f and f' by

$$h(x) = ae^{bx} .$$

The next examples use functions that approximate f , f' and f" locally.

(4) We choose $h(z) = a/(z+b) + c$ and determine a , b , c so that

$$h^{(i)}(x) = f^{(i)}(x) , \quad i = 0,1,2 .$$

We obtain (dropping the argument x):

$$a = -4 f'^3/f''^2$$

$$b = -x - 2 f'/f''$$

$$c = f - 2 f'^2/f'' .$$

Now solving $h(z) = \alpha$ gives the iteration (with $f_n^{(i)} = f^{(i)}(x_n)$)

$$x_{n+1} = x_n - \frac{f_n - \alpha}{f_n'} \cdot \frac{1}{1 - \frac{1}{2}\frac{(f_n - \alpha) f_n''}{f_n'^2}} \tag{6.31}$$

which is Halley's formula to solve $g(x) = f(x) - \alpha = 0$. Recall (see Section 6.1) that the same iteration is obtained if we solve with Newton's method

$$g(x) = \frac{f(x) - \alpha}{\sqrt{f'(x)}} = 0 .$$

If we approximate the inverse function $f^{[-1]}(y)$ locally by $h(y) = a/(y+b) + c$ we get

$$a = -4 f'^3/f''^2$$

$$b = -f + 2 f'^2/f''$$

$$c = x + 2 f'/f'' .$$

Finally putting $x_n = x$ and $x_{n+1} = h(a)$ yields again (6.31). Therefore Halley's formula is obtained by locally approximating f or $f^{[-1]}$ by a hyperbola.

This rational approximation of f has the following property. Suppose we want to solve $f(x) = \alpha$ using Halley's method. We obtain the same iteration (6.31) for

$$g_1(x) = f(x) - \alpha = 0$$

or

$$g_2(x) = \frac{1}{f(x)} - \frac{1}{\alpha} = 0 , \qquad (6.32)$$

i.e., for Halley's method the transformation (6.32) has no effect. But as we saw (6.32) yield another convergence factor for Newton's method.

(5) Euler's method is obtained by approximating f , f' , and f" locally with

$$h(z) = az^2 + bz + c .$$

We get

$$a = f'/2 , \quad b = f' - f''x$$

$$c = f - f' x + \frac{f''}{2} x^2$$

and

$$x_{n+1} = x_n - \frac{(f_n - \alpha)}{f'_n} \cdot G(x_n)$$

with

$$G(x) = 2 \Big/ \left(1 + \sqrt{1 - 2 \frac{(f-\alpha)f''}{f'^2}} \right) .$$

(6) Approximating $f^{[-1]}$ by a parabola

$$h(y) = ay^2 + by + c ,$$

yields

$$a = -\frac{1}{2} \frac{f''}{f'^3}$$

$$b = \frac{1}{f'} + \frac{f''}{f'^3} f$$

$$c = x - \frac{f}{f'} - \frac{1}{2} \frac{f''}{f'^3} f^2 .$$

We obtain the iteration formula for $f(x) - a = 0$

$$x_{n+1} = x_n - \frac{(f_n - \alpha)}{f'_n} \left(1 + \frac{1}{2} \frac{(f_n - \alpha)f''_n}{f'^2_n} \right)$$

(7) Approximating f locally by

$$h(z) = ae^{bz} + c$$

yields

$$a = \frac{f'^2}{f''} \exp - \frac{f''}{f'} x$$

$$b = f''/f'$$

$$c = f - f'^2/f'' .$$

Solving $h(z) = \alpha$ gives the iteration

$$x_{n+1} = x_n + \frac{f'_n}{f''_n} \ln \left(1 - \frac{(f_n - \alpha)f''_n}{f'^2_n} \right)$$

which we can write

$$x_{n+1} = x_n + \frac{f_n - \alpha}{f'_n} G(x_n)$$

with

$$G(x) = -\ln(1 - t(x))/t(x)$$

$$t(x) = (f(x) - a)\, f''(x)/f'^2(x) .$$

From Theorem 6.1 we know that all the methods (4) - (7) are cubically convergent.

6.5 Solving the Secular Equation.

A very good and simple method to solve $f(h) = \alpha^2$ is Reinsch's proposal

$$\lambda_{n+1} = \lambda_n - \frac{f(\lambda_n) - \alpha^2}{f'(\lambda_n)}\, G(\lambda_n) \qquad (6.33)$$

where

$$G(\lambda) = 2\,\frac{\sqrt{f(\lambda)}}{\alpha} \Big/ \left(1 + \frac{\alpha}{\sqrt{f(\lambda)}}\right) .$$

Iteration (6.33) can also be written as

$$\lambda_{n+1} = \lambda_n - \frac{f(\lambda_n)}{f'(\lambda_n)} \cdot 2\left(\frac{\sqrt{f(\lambda_n)}}{\alpha} - 1\right)$$

Reinsch proved in [34] that the sequence λ_n obtained with this iteration converges starting with $\lambda_1 = 0$ monotonically increasing to the solution. This property together with good global convergence makes this iteration very attractive.

If we use another method with starting value $\lambda_1 = 0$, e.g. Halley's iteration, global convergence may not be guaranteed. As an example, consider problem (P3)

$$\min \|\underline{x}\|$$

subject to

$$\|\underline{A}\underline{x} - \underline{b}\| \leq \alpha$$

where $n = m = 6$, $\underline{A}$ Hilbert matrix, $a_{ij} = 1/(i+j-1)$, $\underline{b} = (1,0,0,0,0,0)^T$.

If we start with $\lambda_0 = 0.1$ we get $\lambda_1 = -0.795$ instead of a value bigger than λ_0. The other third order methods involving a term

$$1 - \frac{1}{2}t \quad \text{or} \quad 1 - t \quad , \quad t = \frac{ff''}{f'^2}$$

may also fail if $t > 1$. The quadratic inverse interpolation

$$H(t) = 1 + \frac{1}{2}t \qquad (6.34)$$

yields a correction with the right sign but numerical experiments show that if Halley converges, it converges globally much faster.

An improvement of (6.34) is to use

$$H(t) = \exp(t/2) \quad . \qquad (6.35)$$

However global convergence has to be accelerated by a convergence factor H. Therefore we propose to use

$$\lambda_{n+1} = \lambda_n - \frac{f_n - \alpha^2}{f'_n}\, \exp\left(\frac{1}{2}\,\frac{f_n f''_n}{f'^2_n}\right) \cdot H(\lambda_n) \qquad (6.36)$$

with

$$H(\lambda) = \frac{1}{2}\left(\frac{\sqrt{f(\lambda)}}{\alpha} + \frac{\alpha}{\sqrt{f(\lambda)}}\right) \quad .$$

Since we cannot prove monotonicity or global convergence properties, it is necessary to take measures against too big steps. If the step is too big (usually at the beginning) we use Reinsch's iteration (6.33). For the

above mentioned example with the Hilbert matrix we obtain using (6.36) and $\lambda_0 = 0.1$ the solution of the secular equation $\lambda \cong 7.7 \times 10^{11}$ in four iterations.

To illustrate the behavior of the different methods we computed the only positive solution of the equation

$$f(\lambda) = 1$$

where

$$f(A) = 0.6 + \sum_{i=1}^{20} \frac{2 + 0.8^i}{(\lambda + 0.8^i)^2} \quad .$$

We used ALGOL *W* and double precision (ca. 16 decimal digits) on the IBM 360 of SLAC. The results are displayed in Table 1. Newton's method uses 21 steps to converge to machine precision. Reinsch's variant converges after seven steps. The improved Halley's method needs only four steps.

Table 1

Example for the behaviour of different methods to solve $g(\lambda) = f(\lambda) - \alpha = 0$.

Second order methods $\lambda_{n-1} = F(\lambda_n)$:

Newton $F(\lambda) = \lambda - \frac{g}{g'}$	Reinsch $A - \frac{f}{f'}, 2(\frac{f}{\alpha} - 1)$	Newton for $\ln(f/\alpha) = 0$ $\lambda - \frac{f}{f'}, \ln(\frac{f}{\alpha})$
0.007832098555661240	5.202976431991160	0.063418740364783 4
0.020910105894338 3	7.500022216837 86	0.530745065218083
0.042443876317233 3	9.694555987290 93	1.973582054359389
0.077467170580 8478	10.24226193937 98	4.587394286679 85
0.133829409609385	10.26993613920 28	7.475642842196 32
0.223664677428177	10.27000222396 07	9.527854360359 09
0.365485460269776	10.27000222433 62	10.21358247859 41
0.567031564547462	10.27000222433 62	10.26966838667 10
0.928977169378511	10.27000222433 62	10.27000221262 56
1.449407084177 30	10.27000222433 62	10.27000222433 62
2.227596071203 17	10.27000222433 62	10.27000222433 62
3.360045381985 62	10.27000222433 62	10.27000222433 62
4.925915224578 88	10.27000222433 62	10.27000222433 62
6.871340699770 40	10.27000222433 62	10.27000222433 62
8.793503130706 78	10.27000222433 62	10.27000222433 62
9.972308299401 86	10.27000222433 62	10.27000222433 62
10.25742594 82621	10.27000222433 62	10.27000222433 62
10.26997957457 67	10.27000222433 62	10.27000222433 62
10.27000222426 27	10.27000222433 62	10.27000222433 62

Third order Methods $\lambda_{n+1} = \lambda_n - \frac{g}{g'}, H(t)$ with $t := \frac{g\, g''}{g'^2}$

The following two methods fail for starting values < 6.8:

Ostrowski $H(t) = 1/(1 - t)^{0.5}$	Euler $2/(1 + (1 - 2t)^{0.5})$
6.871340699770 40	6.871340699770 40
11.35785646549 89	8.793503130706 78
10.26159013136 24	10.40624268416 26
10.27000222910 05	10.26995821147 43
10.27000222433 62	10.27000222433 62
10.27000222433 62	10.27000222433 62

. .

Table 1 (cont.)

Halley $1/(1 - t/2)$	inverse interp. $1 + t/2$	$\exp(t/2)$
0	0	0
0.0494307764343190	0.0144232340488059	0.0181702326293991
0.253229265175140	0.0460987990921273	0.0630975341263060
0.986381372252434	0.1134925412666652	0.1695960420329553
3.21756490475624	0.2532995122031215	0.4135192827690038
7.85449586417966	0.5383398900697594	0.9516367427329332
10.2238730991803	1.053105286995540	2.083169645824166
10.2700019998135	2.149570417573700	4.260011924039099
10.2700022243362	4.036423430122260	7.652890892215720
10.2700022243362	6.912724343492040	10.03746804028320
10.2700022243362	9.600920625864520	10.26984359974890
10.2700022243362	10.26349617535140	10.27000222433620
10.2700022243362	10.27000221606120	10.27000222433620
10.2700022243362	10.27000222433620	10.27000222433620
10.2700022243362	10.27000222433620	10.27000222433620

Third order methods with convergence factor $G = \left(\frac{\sqrt{f}}{\alpha} + \frac{\alpha}{\sqrt{f}}\right) / 2$:

$H(t) =$ Halley $G / (1 - t/2)$	$G \exp(t/2)$
0	0
5.07846081438509	1.866758871095732
9.89047170050903	5.513503405770731
10.26991420553395	9.093025644373270
10.27000225433620	10.25032659079000
10.27000222433620	10.27000214107360
10.27000222433620	10.27000222433620

7. Generalizations.

Let A be an $(n \times n)$ matrix, b a given n vector, $\underline{C}$ an $m \times n$ matrix, d a given m vector and γ , α real numbers. We consider the problem

$$F(\underline{x}) = \underline{x}^T \underline{A} \underline{x} + \underline{b}^T \underline{x} + \gamma = \min \tag{7.1}$$

subject to

$$\|\underline{C}\underline{x} - \underline{d}\| = a \quad . \tag{7.2}$$

Using a Lagrange multiplier $\lambda/2$ the solution of (7.1), (7.2) is a stationary point $(\underline{x},\lambda)$ of

$$L(\underline{x},\lambda) = F(\underline{x}) + \frac{\lambda}{2} \{\|\underline{C}\underline{x} - \underline{d}\|^2 - \alpha^2\} \quad . \tag{7.3}$$

$\frac{\partial L}{\partial \underline{x}} = \underline{0}$ and $\frac{\partial L}{\partial \lambda} = 0$ gives

$$\left.\begin{aligned} (\underline{A} + \underline{A}^T + \lambda \underline{C}^T \underline{C})\underline{x} &= - \underline{b} + \lambda \underline{C}^T \underline{d} \\ \|\underline{C}\underline{x} - \underline{d}\|^2 &= \alpha^2 \quad . \end{aligned}\right\} \tag{7.4}$$

Theorem 7.1. Let $(\underline{x}_i,\lambda_i)$, $i = 1,2$, be two solutions of (7.4), then

$$F(\underline{x}_1) - F(\underline{x}_2) = \frac{\lambda_2 - \lambda_1}{4} \|\underline{C}(\underline{x}_1 - \underline{x}_2)\|^2 \quad . \tag{7.5}$$

Proof. This theorem is a generalization of Theorem 2.1 and the proof is very similar. We have

$$(\underline{A} + \underline{A}^T + \lambda_1 \underline{C}^T \underline{C})\underline{x}_1 = - \underline{b} + \lambda_1 \underline{C}^T \underline{d} \tag{7.6}$$

$$(\underline{A} + \underline{A}^T + \lambda_2 \underline{C}^T \underline{C})\underline{x}_2 = - \underline{b} + \lambda_2 \underline{C}^T \underline{d} \tag{7.7}$$

$\underline{x}_1^T$ (7.6) gives

$$2\,\underline{x}_1^T\,\underline{A}\underline{x}_1 + \lambda_1\|\underline{C}\underline{x}_1\|^2 = -\,\underline{x}_1^T\,\underline{b} + \lambda_1\,\underline{x}_1^T\,\underline{C}^T\underline{d} \quad ,$$

or rearranged

$$2(\underline{x}_1^T\,\underline{A}\underline{x}_1 + \underline{x}_1^T\,\underline{b}) = \underline{x}_1^T\,\underline{b} - \lambda_1(\|\underline{C}\underline{x}_1\|^2 - \underline{x}_1^T\,\underline{C}^T\underline{d}) \,. \tag{7.8}$$

Similarly we have

$$2(\underline{x}_2^T\,\underline{A}\underline{x}_2 + \underline{x}_2^T\,\underline{b}) = \underline{x}_2^T\,\underline{b} - \lambda_2(\|\underline{C}\underline{x}_2\|^2 - \underline{x}_2^T\,\underline{C}^T\underline{d}) \,. \tag{7.9}$$

Now $\underline{x}_2^T$ (7.6) - $\underline{x}_1^T$ (7.7) gives

$$(\lambda_1 - A_2)\underline{x}_2^T\,\underline{C}^T\underline{C}\underline{x}_1 = -\underline{x}_2^T\,\underline{b} + \underline{x}_1^T\,\underline{b} + \lambda_1\,\underline{x}_2^T\,\underline{C}^T\underline{d} - \lambda_2\,\underline{x}_1^T\,\underline{C}^T\underline{d}$$

$$\Rightarrow (\underline{x}_1 - \underline{x}_2)^T\underline{b} = (\lambda_1 - \lambda_2)\underline{x}_2^T\,\underline{C}^T\,\underline{C}\underline{x}_1 - d^T(\lambda_1\,\underline{C}\underline{x}_2 - \lambda_2\,\underline{C}\underline{x}_1) \,. \tag{7.10}$$

Subtracting (7.8) - (7.9) and replacing $(\underline{x}_1 - \underline{x}_2)^T b$ by (7.10) we get

$$2(F(\underline{x}_1) - F(\underline{x}_2)) =$$

$$- \lambda_1\{\|\underline{C}\underline{x}_1\|^2 - \underline{x}_1^T\,\underline{C}^T\underline{d} - \underline{x}_2^T\,\underline{C}^T\,\underline{C}\underline{x}_1 + \underline{x}_2^T\,\underline{C}^T\underline{d}\}$$

$$+ \lambda_2\{\|\underline{C}\underline{x}_2\|^2 - \underline{x}_2^T\,\underline{C}^T\underline{d} - \underline{x}_2^T\,\underline{C}^T\,\underline{C}\underline{x}_1 + \underline{x}_1^T\,\underline{C}^T\underline{d}\} \,.$$

Now like in the proof of Theorem 2.1 we know that both { } are equal and therefore we replace them by their arithmetic mean which gives the desired result. Cl

Corollary 7.1. The solution of (7.1), (7.2) is the solution $(\underline{x},\lambda)$ of (7.4) with largest A .

Proof. From (7.5) we have

$$\lambda_2 < A_1 \Rightarrow F(\underline{x}_2) > F(\underline{x}_1) \,. \quad \square$$

Theorem 7.2. Let $(\underline{x}_i,\lambda_i)$, i = 1,2 , be two solutions of (7.4), then

$$(\lambda_1+\lambda_2)(F(\underline{x}_1) - F(\underline{x}_2)) = (\lambda_2 - \lambda_1)(\underline{x}_1 - \underline{x}_2)^T\,\underline{A}(\underline{x}_1 - \underline{x}_2) \,. \tag{7.11}$$

Proof.

$$\lambda_1\,\underline{C}^T\,\underline{C}\underline{x}_1 - \lambda_1\,\underline{C}^T\underline{d} = -(\underline{A}+\underline{A}^T)\underline{x}_1 - \underline{b} \tag{7.12}$$

$$\lambda_2\,\underline{C}^T\,\underline{C}\underline{x}_2 - \lambda_2\,\underline{C}^T\underline{d} = -(\underline{A}+\underline{A}^T)\underline{x}_2 - \underline{b} \,. \tag{7.13}$$

$\lambda_1\underline{x}_1^T$ (7.13) - $\lambda_2\underline{x}_2^T$ (7.12) gives

$$\lambda_1\lambda_2\{(\underline{x}_2 - \underline{x}_1)^T\,\underline{C}^T\underline{d}\} = (\lambda_2 - \lambda_1)\underline{x}_1^T\,(\underline{A}+\underline{A}^T)\underline{x}_2 - (\lambda_1\underline{x}_1 - \lambda_2\underline{x}_2)^T\underline{b} \,. \tag{7.14}$$

$\lambda_1\underline{x}_2^T$ (7.13) - $\lambda_2\underline{x}_1^T$ (7.12) gives

$$\lambda_1\lambda_2\{\|\underline{C}\underline{x}_2\|^2 - \|\underline{C}\underline{x}_1\|^2 + (\underline{x}_1 - \underline{x}_2)\underline{C}^T\underline{d}\}$$

$$= -2\lambda_1\,\underline{x}_2^T\,\underline{A}\underline{x}_2 + 2\lambda_2\,\underline{x}_1^T\,\underline{A}\underline{x}_1 - (\lambda_1\underline{x}_2 - \lambda_2\underline{x}_1)^T\underline{b} \,. \tag{7.15}$$

Observe that

$$0 = \|\underline{C}\underline{x}_2 - \underline{d}\|^2 - \|\underline{C}\underline{x}_1 - \underline{d}\|^2 = \|\underline{C}\underline{x}_2\|^2 - \|\underline{C}\underline{x}_1\|^2 + 2(\underline{x}_1 - \underline{x}_2)^T\,\underline{C}^T\underline{d} \,.$$

So that if we subtract (7.15) - (7.14) we get

$$0 = 2\lambda_2\,\underline{x}_1^T\,\underline{A}\underline{x}_1 - 2\lambda_1\,\underline{x}_2^T\,\underline{A}\underline{x}_2 - (\lambda_1\underline{x}_2 - \lambda_2\underline{x}_1 - \lambda_1\underline{x}_1 + \lambda_2\underline{x}_2)^T\underline{b}$$

$$+ (\lambda_1 - \lambda_2)\,\underline{x}_1^T\,(\underline{A}+\underline{A}^T)\underline{x}_2 \,.$$

Rearranged we have

$$\lambda_1\{2\, x_2^T A x_2 + x_2^T b - x_1^T b - x_1^T (A + A^T) x_2\}$$

$$= \lambda_2\{2\, x_1^T A x_1 + x_1^T b - x_2^T b - x_1^T (A + A^T) x_2\} \quad . \tag{7.16}$$

Now observe that the left hand side of (7.16) is

$$\{\ \} = F(x_2) - F(x_1) + (x_1 - x_2)^T A\, (x_1 - x_2) \quad .$$

Similarly the right hand side simplifies. Therefore

$$\lambda_1\{F(x_2) - F(x_1) + (x_1 - x_2)^T A\, (x_1 - x_2)\}$$

$$= \lambda_2\{F(x_1) - F(x_2) + (x_1 - x_2)^T A\, (x_1 - x_2)\}$$

or rearranged

$$(\lambda_1 + \lambda_2)(F(x_1) - F(x_2)) = (\lambda_1 - \lambda_2)(x_1 - x_2)^T A\, (x_1 - x_2) \quad . \ \square$$

Corollary 7.2. Let (x_i, λ_i), $i = 1,2$, be two solutions of (7.4) with $\lambda_1 \neq \lambda_2$. Then

$$(x_1 - x_2)^T A\, (x_1 - x_2) = -\frac{\lambda_1 + \lambda_2}{4} \|C(x_1 - x_2)\|^2 \quad . \tag{7.17}$$

Proof. We combine the results of Theorems 7.1 and 7.2. $\square$

8. Smoothing of Datas.

Rutishauser proposed in [36] a method to smooth datas. We present here a modified version. Let

$$d_i \ , \quad i = 1,\ldots,n$$

be given datas of a smooth function. However let's assume the d_i are perturbed by measurement errors. We look for a new set of datas

$$x_i \ , \quad i = 1,\ldots,n$$

that does not deviate too much from d_i and that is smoother. If we assume that the d_i arc equidistant then we might want to solve

$$\sum_{i=2}^{n-1} (x_{i+1} - 2x_i + x_{i-1})^2 = \min \tag{8.1}$$

subject to

$$\sum_{i=1}^{n} (x_i - d_i)^2 \leq n\, s^2 \quad . \tag{8.2}$$

Here s^2 is a measure for the variance, the mean deviation we want to allow the new data x_i to differ from d_i :

$$s \approx \sqrt{\frac{\sum_{i=1}^{n} ((x_i - d_i))^2}{n}} \quad .$$

The larger s the more the x_i will be smoothed.

Introducing the $(n-2) \times n$ tridiagonal matrix

$$A = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}$$

and the vectors d and x, equations (8.1) and (8.2) become

$$\left.\begin{aligned} &\|\underline{A}\underline{x}\| = \min \\ &\text{subject to} \\ &\|\underline{x} - \underline{d}\| \leq \alpha := \sqrt{n}\cdot\delta \end{aligned}\right\} \tag{8.3}$$

and we have problem (P1) for $\underline{b} = \underline{0}$ and $\underline{C} = \underline{I}$. The normal equations are

$$(\underline{A}^T\underline{A} + \lambda \underline{I})\underline{x} = \lambda\, \underline{d} \tag{8.4}$$

$$\|\underline{x} - \underline{d}\|^2 = \alpha^2 \quad . \tag{8.5}$$

Instead of solving (8.4), Rutishauser proposed to choose some "smoothing" parameter γ and to solve

$$(\underline{I} + \gamma \underline{A}^T\underline{A})\underline{x} = \underline{d} \tag{8.6}$$

which is (8.4) for $\lambda = 1/\gamma$. The condition number of the matrix in (8.6) is 16γ. For large n one may have to choose γ large which leads to numerical problems [36]. However large γ or small λ may be meaningful: for $\gamma \to \infty$ ($\lambda \to 0$) the new values $\underline{x}$ lie on a straight line, i.e., are values of a linear function. Of course one would like to have as limit the linear regression to the datas d.

If we make the change of variables in (8.3)

$$\begin{aligned} \underline{w} &:= \underline{x} - \underline{d} \\ b &:= -\underline{A}\underline{d} \end{aligned} \tag{8.7}$$

then our problem becomes

$$\left.\begin{aligned} &\|\underline{A}\underline{w} - \underline{b}\| = \min \\ &\text{subject to} \\ &\|\underline{w}\| \leq \alpha\,. \end{aligned}\right\} \tag{8.8}$$

Problem (8.8) leads to a relaxed least squares problem

$$\left.\begin{aligned} &(\underline{A}^T\underline{A} + \lambda \underline{I})\underline{w} = \underline{A}^T\underline{b} \\ &\|\underline{w}\| = \alpha \end{aligned}\right\} \tag{8.9}$$

We can use the dual equations for (8.9) (see Section 3):

$$(\underline{A}\underline{A}^T + \lambda \underline{I})\underline{z} = -\underline{b} = \underline{A}\underline{d} \tag{8.10}$$

$$\|\underline{A}^T\underline{z}\| = \alpha \tag{8.11}$$

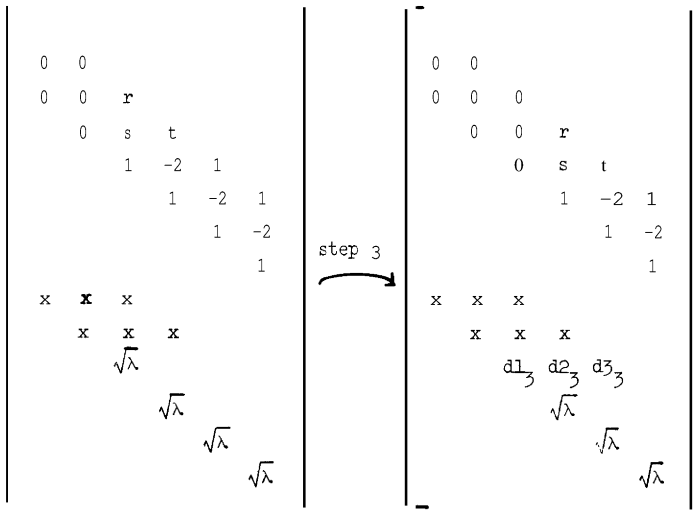
with

$$w = -\underline{A}^T\underline{z} \quad . \tag{8.12}$$

Observe that now $\lambda \to 0$ causes no problems since $\underline{A}\underline{A}^T$ is not singular. Furthermore since $\underline{b} = -\underline{A}\underline{d}$ we can write (8.10) as

$$\begin{pmatrix} \underline{A}^T \\ \sqrt{\lambda}\,\underline{I} \end{pmatrix}\underline{z} \approx \begin{pmatrix} \underline{d} \\ \underline{0} \end{pmatrix} \quad . \tag{8.13}$$

A is tridiagonal, therefore we shall use the algorithm described in Section 5.1 to solve for a given $A \geq 0$ the least squares problem (8.13). For n = 8, e.g. the remaining matrix before and after the third step is:



The following algorithm transforms

$$\underline{U}^T\begin{pmatrix}\underline{A}^T\\ \sqrt{\lambda}\,\underline{I}\end{pmatrix} = \begin{pmatrix}\underline{0}\\ \underline{R}\end{pmatrix} \quad \text{and} \quad \underline{U}^T\begin{pmatrix}\underline{d}\\ \underline{0}\end{pmatrix} = \begin{pmatrix}\underline{c}\\ \underline{y}\end{pmatrix}$$

where U is orthogonal and R upper triangular and **tridiagonal**:

$$R = \begin{vmatrix} dl_1 & d2_1 & d3_1 & & & \\ & dl_2 & d2_2 & d3_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & d3_{n-4} \\ & & & & \ddots & d2_{n-3} \\ & & & & & dl_{n-2} \end{vmatrix}$$

The matrix A is not stored, we work only with three simple variables r , s , and t . We use the notation

$$\begin{pmatrix} 0 & a'_2 & a'_3 \\ b'_1 & b'_2 & b'_3 \end{pmatrix} := \underline{G}\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \qquad (8.14)$$

where G is a Givens matrix

$$\underline{G} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

and a has been chosen so that the zero in the left hand side of (8.14) appears.

```
begin
    comment zeroing columns 1 to n-4;
    r := t := 1; s := -2; c_1 := d_1; c_2 := d_2;
    for i := 1 step 1 until n-4 do
    begin
```

$$\begin{pmatrix} 0 & c_i \\ dl_i & y_i \end{pmatrix} := \underline{G}_1\begin{pmatrix} r & c_i \\ \sqrt{\lambda} & 0 \end{pmatrix};$$

$$\begin{pmatrix} 0 & r & c_{i+1} \\ dl_i & d2_i & y_i \end{pmatrix} := \underline{G}_2\begin{pmatrix} s & t & c_{i+1} \\ dl_i & 0 & y_i \end{pmatrix};$$

$$\begin{pmatrix} 0 & s & t & c_{i+2} \\ dl_i & d2_i & d3_i & y_i \end{pmatrix} := \underline{G}_3\begin{pmatrix} 1 & -2 & 1 & d_{i+2} \\ dl_i & d2_i & 0 & y_i \end{pmatrix};$$

```
    end;
```

comment column n-3 ;

$$\begin{pmatrix} 0 & c_{n-3} \\ dl_{n-3} & y_{n-3} \end{pmatrix} := G_1 \begin{pmatrix} r & c_{n-3} \\ \sqrt{\lambda} & 0 \end{pmatrix} ;$$

$$\begin{pmatrix} 0 & r & c_{n-2} \\ dl_{n-3} & d2_{n-3} & y_{n-3} \end{pmatrix} := G_2 \begin{pmatrix} s & t & 'n-2 \\ dl_{n-3} & 0 & y_{n-3} \end{pmatrix} ;$$

$$\begin{pmatrix} 0 & s & c_{n-1} \\ dl_{n-3} & d2_{n-3} & y_{n-3} \end{pmatrix} := G_3 \begin{pmatrix} 1 & -2 & d_{n-1} \\ dl_{n-3} & d2_{n-3} & y_{n-3} \end{pmatrix} ;$$

comment column n-2;

$$\begin{pmatrix} 0 & c_{n-2} \\ dl_{n-2} & y_{n-2} \end{pmatrix} := G_1 \begin{pmatrix} r & c_{n-2} \\ \sqrt{\lambda} & 0 \end{pmatrix} ;$$

$$\begin{pmatrix} 0 & c_{n-1} \\ dl_{n-2} & y_{n-2} \end{pmatrix} := G_2 \begin{pmatrix} s & c_{n-1} \\ dl_{n-2} & y_{n-2} \end{pmatrix} ;$$

$$\begin{pmatrix} 0 & c_n \\ dl_{n-2} & y_{n-2} \end{pmatrix} := G_3 \begin{pmatrix} 1 & d_n \\ dl_{n-2} & y_{n-2} \end{pmatrix} ;$$

end.

To solve the problem we need the derivatives of $f(\lambda)$ defined by

$$(\underline{A}\underline{A}^T + \lambda \underline{I})\underline{z}(\lambda) = \underline{A}\ \underline{d}$$

$$f(\lambda) = \|\underline{A}^T \underline{z}(\lambda)\|^2 .$$

We cannot apply all the results of Section 5.4 since f is not of the ssme form. We have

$$(\underline{A}\underline{A}^T + \underline{A}\ \underline{I})z^{(k)} = - z^{(k-1)} , \quad k \geq 1 .$$

We shall use Reinsch's proposal and therefore we need only f' :

$$f'(A) = 2\ \underline{z}'^T\ \underline{A}\ \underline{A}^T\ \underline{z}$$

$$= - 2\ \underline{z}^T(\underline{z} + \lambda\ z') .$$

An ALGOL W procedure **smooth(n,δ,d,x)** is given at the end of this section. The following example has been computed with this procedure using single precision (ca. 6 decimal digits) on the SLAC computer (IBM 360).

Example.

we choose n = 30 and

$$d_i = \sqrt{i} + 0.2 \sin(i) , \quad i = 1,\ldots,30 .$$

For this example we have f(0) = 1.825814 , therefore if

$$\delta > \sqrt{\frac{f(0)}{n}} = 0.246699$$

the smoothed value x_i are the same as obtained by linear regression.

On the other hand if $\delta = 10^{-4}$ we have $\|\underline{x} - \underline{d}\| = 5.4 \times 10^{-4}$ and the x_i differ only in the fourth decimal from the d_i . If we interpret the d_i as perturbed value of $\sqrt{i}$ then because the mean of $|0.2 \sin(i)|$ is approximately

$$0.2\,\frac{1}{\pi}\int_0^{\pi}\sin(x)\,dx \qquad \frac{0.4}{\pi} \approx 1.3$$

we expect the best smoothing for $\delta \approx 1.3$ which indeed can be seen clearly in Table 2.

Table 2 : Smoothing of $d_i = \sqrt{i} + 0.2 \sin(i)$, $i = 1, \ldots, 30$.

δ =	$\geq$ 0.2467	0.2466	0.2	0.17	0.15	0.13	0.12
	1.722253	1.722027	1.603430	1.507798	1.414468	1.261987	1.243047
	1.861272	1.861084	1.763119	1.684705	1.609509	1.496164	1.483359
	2.000287	2.000137	1.922686	1.861354	1.804054	1.727389	1.717025
	2.139301	2.139186	2.081965	2.03'1419	1.997584	1.955855	1.947423
	2.270316	2.278234	2.240736	2.212497	2.189486	2.182'792	2.181764
	2.417336	2.417287	2.398729	2.386040	2.378851	2.406053	2.418491
	2.556347	2.556335	2.555611	2.557375	2.564482	2.61912;	2.643821
	2.695338	2.695371	2.711038	2.725835	2.745208	2.815103	2.841743
	2.034349	2.834425	2.864746	2.890921	2.920292	2.992062	3.003115
	2.973354	2.973438	3.016541	3.052362	3.089548	3.154735	3.155210
	3.112376	3.112476	3.166302	3.210031	3.253120	3.310709	3.301294
	3.251370	3.251490	3.313931	3.363794	3.411095	3.464372	3.457992
	3.390381	3.390501	3.459267	3.513468	3.563265	3.613997	3.616866
	3.529382	3.529527	3.602212	3.658838	3.70331'1	3.754458	3.762860
	3.668403	3.668556	3.742742	3.799843	3.849200	3.883019	3.886796
	3.307439	3.807585	3.880861	3.936624	3.983325	4.002785	3.995223
	3.946456	3.946573	4.016701	4.069475	4.112409	4.120646	4.105030
	4.085469	4.085618	4.150337	4.196711	4.237081	4.241594	4.228400
	4.224488	4.22460'1	4.282032	4.324522	4.357614	4.364625	4.362045
	4.363507	4.363605	4.411738	4.447031	4.473988	4.404037	4.4905'13
	4.502520	4.502613	4.539614	4.566405	4.586239	'I. 594892	4.600982
	4.641560	4.641610	4.665841	4.682961	4.694780	4.697631	4.694865
	4.780583	4.730605	4.790689	4.797174	4.800344	4.797594	4.787060
	4.919610	4.919595	4.914412	4.400512	4.903647	4.899828	4.891846
	5.058637	5.058589	5.037205	5.020307	5.005067	5.003985	5.001230
	5.137662	5.197586	5.159234	5.129760	5.104605	5.103989	5.123779
	5.336688	5.336576	5.280628	5.238020	5.202195	5.192799	5.216838
	5.475712	5.475564	5.401550	5.345339	5.298068	5.268023	5.281142
	5.614742	5.614557	5.522194	5.452085	5.392826	5.333412	5.324587
	- 5.753768	-5.753551	5.642737	5.558625	5.487171	5.395162	5.3601306
$\|\| \underline{x} - \underline{d} \|\|$	1.351226	1.350682	1.095428	0.9311236	0.8215933	0.7120381	0.6572663
$\left[\sum (\Delta^2 x_i)^2\right]^{1/2}$	$6.67\ 10^{-5}$	$1.11\ 10^{-4}$	0.008449	0.0150982	0.021434	0.0445737	0.0796041
$\left[\sum (x_i - \sqrt{i})^2\right]^{1/2}$	1.205420	1.204810	0.9014646	0.6828826	0.5084146	0.306330	0.3031789
# of iteration	0	1	5	5	7		
$\lambda = f^{-1}(n\delta^2)$	0	$3.83\ 10^{-7}$	$2.79\ 10^{-4}$	$7.60\ 10^{-4}$	$2.01\ 10^{-3}$	$3.15\ 10^{-2}$	$8.89\ 10^{-2}$

Table 2 (cont.)

δ =	0.1	0.07	0.05	0.01	0.001	0.0001	0 (d_1)
	1.238009	1.232636	1.222315	1.183728	1.169964	1.168462	1.168294
	1.483301	1.497992	1.514537	1.571436	1.593233	1.595784	1.596072
	1.710183	1.713835	1.723948	1.755995	1.760237	1.760216	1.760274
	1.930025	1.906127	1.891484	1.855504	1.849538	1.848726	1.848639
	2.167420	2.136563	2.111879	2.056374	2.045358	2.044389	2.044283
	2.425476	2.422603	2.415583	2.397161	2.393924	2.393637	2.393606
	2.674776	2.710691	2.731615	2.769294	2.776435	2.777078	2.777148
	2.877497	2.924973	2.956028	3.011403	3.025218	3.026191	3.026293
	3.022841	3.040714	3.053176	3.077481	3.081974	3.082378	3.082423
	3.139317	3.111163	3.092867	3.060061	3.054076	3.053531	3.053473
	3.271142	3.221653	3.188797	3.128712	3.111723	3.117362	3.116626
	3.439886	3.413122	3.395536	3.363274	3.357374	3.356843	3.356786
	3.626307	3.645697	3.659373	3.684522	3.689124	3.689538	3.689534
	3.782226	3.836155	3.866444	3.927829	3.938693	3.939671	3.939773
	3.904166	3.935042	3.956223	3.995197	4.002328	4.002970	4.003040
	3.986406	3.972642	3.963206	3.945898	3.942735	3.942450	3.942419
	4.077334	4.031558	4.000144	3.942432	3.931878	3.930929	3.930425
	4.206725	4.171137	4.146597	4.101509	4.093266	4.092524	4.092443
	4.515091	4.559080	4.588960	4.643706	4.653724	4.654625	4.654724
	4.622007	4.662054	4.689597	4.739809	4.749388	4.749315	4.749906
	4.690324	4.689064	4.669136	4.688752	4.688654	4.688645	4.688644
	4.757719	4.714632	4.687192	4.636802	4.627513	4.626677	4.626586
	5.004077	4.990499	4.983374	4.975096	4.973674	4.973543	4.973529
	5.150356	5.180526	5.200182	5.242207	5.250694	5.251443	5.251531
	5.256895	5.303260	5.329032	5.375997	5.386373	5.387322	5.387426
	5.307034	5.333555	5.342944	5.345037	5.345425	5.345654	5.345683
	5.318588	5.311031	5.302564	5.268337	5.254272	5.252623	5.252437
	5.319581	5.284638	5.272645	5.273405	5.278864	5.279541	5.279616
$\equiv \|x - \bar{a}\|$	0.547718	0.383404	0.273859	0.054773	$5.47\ 10^{-3}$	$5.47\ 10^{-4}$	0
$\left[\Sigma\ (\Delta^2 x_i)^2\right]^{1/2}$	0.168388	$0.^{0}$ 316081	0.420681	$0.^{0}$ 647135	0.703808	0.709727	0.710384
$\left[\Sigma\ (x_i - (f^{-1})^2\right]^{1/2}$	0.358399	$0.^{0}$ 479169	0.567851	$0.^{0}$ 746946	0.784216	0.787952	**0.7883406**
# of iteration	6	5	5	5	4	3	3
$\lambda = f^{-1}(n\delta^2)$	0.264058	$0.^{0}$ 769474	1.500484	13.16168	153.4504	1562.654	

```
PROCEDURE SMOOTH(INTEGER VALUE N;
                 REAL VALUE DELTA ;
                 REAL ARRAY D,X(*)) ;
BEGIN
   REAL G,FO,FA,F1,F2,LAMB,LAMBN,WLAMB,R,S,T,ALPHA,ALPHA2,CO,SI,H ;
   REAL ARRAY  C,AZ,AZS(1::N) ;
   REAL ARRAY D1,Y,Z,ZS(1::N-2) ;
   REAL ARRAY D2(1::N-3); REAL ARRAY D3(1::N-4) ;
   INTEGER I ;
   PROCEDURE  ATZ(REAL ARRAY Z,Y(*)) ;
   EEGIN INTEGER I ;
      Y(1) :=  Z(1);
      Y(2) :=  -2*Z(1) + Z(2) ;
      FOR I := 3 STEP 1 UNTIL N-2 DO
         Y(I) := Z(I-2) -2*Z(I-1) + Z(I) ;
      Y(N-1) :=Z(N-3) -2*Z(N-2)  ;
      Y(N) := Z(N-2) ;
   END ATZ ;
   PROCEDURE  BACK(INTEGER VALUE N;REAL ARRAY A,E,C,D,X(*)) ;
   EEGIN INTEGER I ;
      X(N) := D(N)/A(N) ;
      X(N-1) := (D(N-1) - X(N)*B(N-1))/A(N-1) ;
      FOR I := N-2 STEP -1 UNTIL 1 DO
         X(I) := (D(I) - X(I+1)*B(I) - X(I+2)*C(I))/A(I) ;
   END BACK ;
   PROCEDURE VORW(INTEGER VALUE N;REAL ARRAY A,B,C,D,X(*)) ;
   BEGIN INTEGER I ;
      X(1) := D(1)/A(1) ;
      X(2) := (D(2) - X(1)*B(1))/A(2) ;
      FOR I := 3 STEP 1 UNTIL N DO
         X(I) := (D(I) - X(I-1)*B(I-1) - X(I-2)*C(I-2))/A(I) ;
   END VORW ;
   PROCEDURE  ROT(REAL VALUE A;REAL VALUE B);
   BEGIN  REAL T ;
      IF  B = 0  THEN
         BEGIN CO := 0 ; SI := 1 END
      ELSE
         EEGIN
            T := -A/B ; CO := 1/SQRT(1 + T**2) ;
            SI := T*CO ;
         END ;
   END  HOT ;
   REAL PROCEDURE INPROD(INTEGER VALUE N;  REAL ARRAY X,Y(*)) ;
   BEGIN INTEGER I ; REAL S ;
      S := 0 ;
      FOR I :=  1 STEP 1 UNTIL N DO S := S + X(I)*Y(I) ;
      S
   END INPROD ;
   ALPHA := SQRT(N)*DELTA; ALPHA2 := ALPHA**2 ;
   LAMB := 0 ;
```

```
:    WLAMB := SQRT(LAMB) ;
     COMMENT COLUMNS 1 TO N-4 ;
     R := T := 1 ; S :=-2 ; C(1) := D(1); C(2) := D(2) ;
     FOR I := 1 STEP 1 UNTIL N-4 DO
     BEGIN
        ROT(R,WLAMB) ;
        D1(I) := -SI*R + CO*WLAMB ;
        Y(I) := -SI*C(I) ; C(I) := CO*C(I) ;
        ROT(S,D1(I)) ;
        D1(I) := -SI*S + CO*D1(I) ;
        R := CO*T ; D2(I) := -SI*T ;
        H := CO*C(I+1)+SI*Y(I); Y(I):=-SI*C(I+1)+CO*Y(I) ;
        C(I+1) := H ;
        ROT(1,D1(I)) ;
        D1(I) := -SI + CO*D1(I) ;
        S := -2"CO + SI*D2(I) ; D2(I) := 2*SI + CO*D2(I) ;
        T := CO ; D3(I) := -SI ;
        H := CO*D(I+2) + SI*Y(I) ;
        Y(I) := -SI*D(I+2) + CO*Y(I) ; C(I+2) := H ;
     END ;
     FOR I := N-3,N-2 DO
     BEGIN
        COMMENT COLUMNS N-3 AND N-2 ;
        ROT(R,WLAMB) ;
        D1(I) := -SI*R + CO*WLAMB ;
        Y(I) := -SI*C(I) ; C(I) := CO*C(I) ;
        ROT(S,D1(I)); ;
        D1(I) := -SI*S + CO*D1(I) ;
        IF I =N-3 THEN
        BEGIN
           R := CO*T ; D2(I) := -SI*T ;
        END ;
        H := CO*C(I+1) + SI*Y(I) ;
        Y(I) := -SI*C(I+1) + CO*Y(I) ; C(I+1) := H ;
        ROT(1,D1(I) );
        D1(I) := -SI + CO*D1(I) ;
        IF I = N-3 THEN
        BEGIN
           S := -2'CO + SI*D2(I); D2(I):= 2*SI + CO*D2(I) ;
        END ;
        C(I+2) := CO*D(I+2) + SI*Y(I) ;
        Y(I) := -SI*D(I+2) + CO*Y(I) ;
     END I ;
     IF LAMB ▫ 0 THEN FA := 2*INPROD(N,C,C) ;

     BACK(N-2,D1,D2,D3,Y,Z) ;
     ATZ(Z,AZ) ;
     FO := INPROD(N,AZ,AZ) ;
     VORW(N-2,D1,D2,D3,Z,ZS) ;
     COMMENT ZS IS -Z' ;
     BACK(N-2,D1,D2,D3,ZS,ZS) ;
     ATZ(ZS,AZS) ;
```



```
     F1 := -2*INPROD(N,AZS,AZ) ;
     LAMBN := LAMB - FO/F1*2*(SQRT(FO)/ALPHA - 1) ;
     IF (FA <= FO) OR (FO < ALPHA2) OR (LAMBN <= LAMB)
     THEN GOTO FIN ;
     FA := FO ; LAMB := LAMBN ;
     GOTO IT ;
FIN:
     FOR I := 1 STEP 1 UNTIL N DO
        X(I) := D(I) - AZ(I) ;
  END SMOOTH ;
```

References

[1] Bjork, A., "Iterative Refinement of Linear Least Squares Solutions I," BIT 7 (1967), 257-278.

[2] Bjork, A., "Solving Linear Least Squares Problems by Gram-Schmidt Orthogonalization," BIT 7 (1967), 1-21.

[3] Brown, G. H. Jr., "On Halley's Variation of Newton's Method," American Mathematical Monthly 84,9 (1977).

[4] Buchh, J. R. and Rose, D. J., Sparse Matrix Computations, Academic Press, (1976).

[5] Chan, T. F. C., "On Computing the Singular Value Decomposition," Stanford Computer Science Department Report STAN-CS-77-588, (1977).

[6] Davies, M. and Dawson, B., "On Global Convergence of Halley's Iteration Formula," Numer. Math. 24(1975).

[7] Ehrmann, H., "Konstruktion und Durchfuehrung von Iterationsverfahren hoeherer Ordnung," Arch. Rational Mech. Anal. 4 (1959).

[8] Elden, L., "Algorithms for the Regularization of Ill-Conditioned Least Squares Problems," BIT 17 (1977), 134-145.

[9] Elden, L., "Numerical Analysis of Regularization and Constrained Least Squares Methods," Thesis No. 21, Linkopping University (1977).

[10] Forsythe, G. E. and Golub, G. H., "On the Stat. Values of a Second Degree Polynomial on the Unit Sphere," J. Soc. Indust. Appl. Math. 13 (1965).

[11] Forsythe, G. E., Malcolm, M. A. and Moler, C. B., Computer Methods for Mathematical Computations, Prentice Hall, 147.

[12] Gander, W., "How to Apply the Dual Problem to Solve a Certain Constrained Minimum Norm Problem," SIAM Spring Meeting Madison 1978.

[13] Gander, W., Molinari, L. and Svecowa, H., Numerische Prozeduren . . . . ISNM 33, Birkhaeuser Verlag, 1977.

[14] Golub, G. H. and Kahan, W., "Calculating the Singular Values and Pseudo-Inverse of a Matrix," SIAM J. Numer. Anal. 2, 2 (1965).

[15] Golub, G. H., Klema, V. and Stewart, G. W., "Rank Degeneracy and Least Squares Problems," Stanford Computer Science Department Report STAN-CS-76-559 (1976).

[16] Golub, G. H., "Some Modified Matrix Eigenvalue Problems," SIAM Review 15, 2 (1973).

[17] Golub, G. H., Heath, M. and Wahba, G., "Generalized Cross-Validation," Stanford Computer Science Department Report STAN-CS-77-622 (1977).

[18] Golub, G. H. and Luk, F. T., "Singular Value Decomposition: Applications and Computations," Internal Report Serra House Stanford University Computer Science Department (1978).

[19] Hansen, E. and Patrick, M., "A Family of Root Finding Methods," Numer. Math. 27 (1977), 257-269.

[20] Hanson, R. J. and Phillips, J. L., "An Adaptive Numerical Method for Fredholm Integral Equations," Numer. Math. 24 (1975), 291-307.

[21] Henrici, P., Applied and Computational Complex Analysis, Wiley, 1974.

[22] Henrici, P., Elements of Numerical Analysis, Wiley, 1964.

[23] Kahan, W., Manuscript in Box 5 G of the G. Forsythe archive, Stanford Main Library.

[24] Kahan, W., "Why Use Tangents When Secants Will Do," Gatlinburg Conference 1977, Asilomar.

[25] Kalman, R. E., "Algebraic Aspects of the Generalized Inverse ...." Generalized Inverses and Applications, Academic Press, (1976).

[26] Lawson, Ch. L. and Hanson, R. J., Solving Least Squares Problems, Prentice Hall, 1974.

[27] Marquart, D. W., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," SIAM 11, 2 (1963).

[28] Marti, J. T., 'Minimum Norm Solutions of Fredholm Integral Equations of the First Kind," Report 78-01 Sem. f. angew. Math. ETHZ, (1978).

[29] Moré, J. J., "The Levenberg-Marquart Algorithm: Implementation and Theory," Dundee Conference on Numerical Analysis (1977).

[30] Ostrowski, A. M., Solution of Equations and Systems of Equations, Academic Press, (1973).

[31] Paige, C. C., "Bidiagonalization of Matrices and Solution of Linear Equations," SIAM J. Numer. Anal. 11, 1 (1974).

[32] Peters, G. and Wilkinson, J. H., " Ax = λBx and the Generalized Eigenproblem,' SIAM J. Numer. Anal. 7, 4 (1970).

[33] Reinsch, Chr. H.. "Smoothing by Spline Functions," Numer. Math. 10 (1967), 177-183.

[34] Reinsch, Chr. H., "Smoothing by Spline Functions. II, " Numer. Math. 16 (1971), 451-454.

[35] Rutishauser, H., "Once Again: The Least Square Problem," Linear Algebra and Its Applications 1 (1968), 479-488.

[36] Rutishauser, H., "Vorlesungen ueber numerische Mathematik, hrsg. von M. Gutknecht," Band I, II, Birkhaeuser Verlag, 1976.

[37] Schek, H.-J. and Eggensperger, R., "Least Squares - Loesungen und Daempfung bei unterbestimmten Gleichungssystemen," Computing 19 (1977).

[38] Schwarz, H. R., Rutishauser, H. and Stiefel, E., Matrizen-Numerik, Teubner Verlag, 1968.

[39] Spjøtvoll, E., "A Note on a Theorem of Forsythe and Golub," SIAM J. Appl. Math. 23, 3 (142).

[40] Stewart, G. W., "On the Continuity of the Generalized Inverse," SIAM J. Appl. Math. 17, 1 (1969).

[41] Stewart, G. W., "On the Numerical Properties of an Iteration for Computing the Generalized Inverse," CNA 12, Austin, Texas (1971).

[42] Tikhonov, A. N., Solution of Ill Posed Problems, Wiley, 1977.

[43] Traub, J. F., Iterative Methods for Solution of Equations, Prentice Hall, 1964.

[44] Van Loan, Ch. "Lectures in Least Squares," Technical Report TR 76-279, Cornell University (1976).

[45] Van Loan, Ch. F., "Generalizing the Singular Value Decomposition," SIAM J. Numer. Anal. 13, 1 (1976).

[46] Varah, J. M., "On the Numerical Solution of Ill-Conditioned Linear Systems," SIAM J. Num. Anal. 10 (1973).

[47] Varah, J. M., "A Practical Examination of Num. Methods for Ill Posed Problems," Technical Report 76-08, University of British Columbia, 1976.

[48] Wilkinson, J. H. and Reinsch, C., Linear Algebra, Springer, 1971.