

Tilburg University

A Newton-like method for error analysis

Paardekooper, M.H.C.

Publication date:
1981

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

Paardekooper, M. H. C. (1981). *A Newton-like method for error analysis: Applied to linear continuous systems and eigenproblems*. (Research memorandum / Tilburg University, Department of Economics; Vol. FEW 105). Unknown Publisher.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

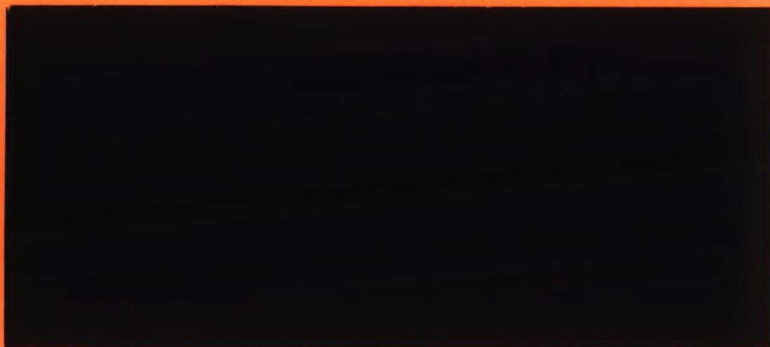
CBM
R



7626
1981
105



subfaculteit der econometrie

RESEARCH MEMORANDUM



Bestemming 	TIJDSCHRIFTENBUREAU BIBLIOTHEEK KATHOLIEKE HOGESCHOOL TILBURG	Nr. 
--	---	--

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS

Postbus 90135 - 5000 LE Tilburg
Netherlands



 K.U.B.
BIBLIOTHEEK
TILBURG

SUBFACULTEIT DER ECONOMETRIE

A NEWTON-LIKE METHOD FOR ERROR ANALYSIS

APPLIED TO LINEAR CONTINUOUS SYSTEMS AND EIGEN-
PROBLEMS

M.H.C. Paardekooper

december 1981

A NEWTON-LIKE METHOD FOR ERROR ANALYSIS

APPLIED TO LINEAR CONTINUOUS SYSTEMS AND EIGENPROBLEMS

ABSTRACT

This paper discusses bounds for the distance of a point a in Hilbert space X to a manifold $S = \{x \in X | F(x) = 0\}$, where f is a Frechet differentiable mapping into a Hilbert space Y . The Newton-Kantorovich method handled in the normal space at a of the manifold $\tilde{S} = \{x \in X | F(x) = F(a)\}$, gives a realistic upperbound of this distance. In the usual Kantorovich conditions the Lipschitz continuity of Df effects S to be locally in a set with positive distance to a ; this leads to a lower bound of the distance $d(a,S)$. This approach leads to a method to distribute, in an "optimal" way the errors in a linear dynamical system among input, output and structural variables. A posteriori error analysis in eigenproblems illustrates also this approach.

CONTENTS

1. Introduction	1
2. Bounds for the distance of a manifold in a Hilbert space to a nearby point.	4
3. Distribution of errors in a linear continuous system.	13
4. A posteriori error bounds for an eigenpair and numerical results.	24
5. References.	33

1. INTRODUCTION

In applied mathematics it is of frequent occurrence to have only an approximate solution of an equation. Only approximately the data satisfy the model equations, despite accuracy the numerical solutions have been applied with errors, the structural parameters supposed to be constant are as well approximately known,...

For this reason it is important to have at disposal a manageable technique to determine sharp bounds for these errors. Both upper and lower bounds for the errors are of importance in the evaluation and diagnostics of models or to assess a numerical algorithm.

In the late forties Kantorovich proved the convergence of Newton's method for solving equations (, without assuming previously the existence of a solution,) under meanwhile standard assumptions [9], [13]. As an impressive result he also obtained an upperbound of the distance of starting value x_0 to the locally unique solution x_∞ , being the limit of the Newton sequence. Lower bounds of this distance are given by Gragg and Tapia [8]. In the Newton-Kantorovich theory one considers the convergence of the Newton sequence to a zero of the sufficiently smooth function

$$f : X \supset W \rightarrow Y \quad (1.1)$$

where W is some ball around $a = x_0$ in a Banach space X and the inverse of derivative $Df(0)$ maps Banach space Y onto X . At the moment we remind and emphasize the Kantorovich-Gragg-Tapia bound for $x_\infty - x_0$:

$$m \|Df(x_0)^{-1} f(x_0)\| \leq \|x_\infty - x_0\| \leq M \|Df(x_0)^{-1} f(x_0)\|, \quad (1.2)$$

where, according to [13] $M = (1 - \sqrt{1-2\kappa})/\kappa$ and according to [8] $m = (\sqrt{1+2\kappa} - 1)/\kappa$. Here $\kappa = L\mu^{-1} \|Df(x_0)^{-1} f(x_0)\| < \frac{1}{2}$ where L is a Lipschitz constant of Df and $\|Df(x_0)^{-1}\| \leq \frac{1}{\mu}$.

In section two of this memorandum we discuss a generalization of the Newton-Kantorovich theorem. We consider a function f as given in (1.1), but now Frechet derivative $Df(0): X \rightarrow Y$ is surjective. A lower and an upper bound is derived for the distance $d(a,S)$ of $a \in W$ to the manifold

$S = \{x \in W \mid f(x) = 0\}$. As concerns f we assume derivative $Df: W \rightarrow L(X, Y)$ to be Lipschitz continuous. Let be φ the restriction of f to the normal space N_2 at $a \in W$ to the level surface $\{x \in W \mid f(x) = f(a)\}$.

Provided φ satisfies the usual Kantorovich conditions, the intersection of W and N_2 contains a zero, say z , of $f: z \in S$. So $\|a - z\|$ is an upperbound of $d(a, S)$. Now the Lipschitz continuity of Df has as consequence that S around z fluctuates inside a set that can be described by a simple inequality; this inequality gives a lowerbound of $d(a, S)$.

In practical situations many times the problem under consideration rises in the following way. Assume in some model, with given parameters $a \in \mathcal{P}$, some variables $v \in \mathcal{V}$ have to satisfy the equation $G(y; a) = 0$. As a result of errors $G(v, a) \neq 0$. Then the question rises to find bounds for the perturbation $e \in \mathcal{P}$ and $d \in \mathcal{V}$ such that $f(d, e) = 0$, where

$$f(d, e) = G(v + d; a + e).$$

In previous papers [14], [15] we have found realistic, and rather usable upperbounds for these feasible perturbations $(d, e) \in \mathcal{V} \times \mathcal{V}$ in the least squares problem and in the discrete linear model.

In section three we analyse the magnitude of sufficient perturbations $(\alpha, \beta, \rho, \xi, \eta)$ in the continuous linear model

$$f(\alpha, \beta, \rho, \xi, \eta)(t) := \dot{y} + \dot{\eta} - (a + \alpha(t))(y + \eta(t)) -$$

$$(b + \beta(t))(x + \xi(t)) - \rho(t) = 0 \tag{1.3}$$

with given $a, b \in \mathbb{R}$, $x \in C[0, 1]$ and $y \in C^1[0, 1]$.

The non-linearity of f makes it necessary to use supremum norms and the contraction theorem in, say, $\text{im}(Df(0, 0, 0, 0, 0))^*$.

Hence only an upperbound of $d(0, S)$ has been obtained; for the rest, that upperbound is sharp for small $r := f(0, 0, 0, 0, 0)$.

In section four we consider the eigenproblem as an area of application of the mentioned approach. There we describe, in addition to the error bounds, also a method for the improvement of an approximated eigenpair (λ, x) : a zero (δ, h) of the function $A(x + h) - (\lambda + \delta)(x + h)$ is a solution of an underdetermined system [2], [16]. But given that (λ, x) also the question

rises to construct an optimal perturbation (H, h) such that $(A + H)(x + h) - \lambda(x + h) = 0$. Numerical results for this problem demonstrate that if λ is an ill considered eigenvalue than a very small H suffices to fulfill the equation $(A + H)(x + h) = \lambda(x + h)$.

2. BOUNDS FOR THE DISTANCE OF A MANIFOLD IN A HILBERT SPACE TO A NEARBY POINT.

In this section we describe a method to determine an upper and a lower bound for the distance of a given point a in a Hilbert space X to a manifold S .

Let S be a manifold imbedded in X , defined as the preimage of zero in Y of a surjective mapping f of X into Hilbert space Y :

$$S := \{x \in X \mid f(x) = 0\} \quad (2.1)$$

In X the distance of a to S is defined by

$$d(a, S) = \inf\{\|x - a\| \mid x \in S\}.$$

The bounds for $d(a, S)$ are derived under the assumptions that f is a mapping with Lipschitz continuous derivatives in some ball around a and that $f(a)$, $Df(a)$ and this Lipschitz constant for Df fit into the wellknown Kantorovich conditions. So the point a is near to S as can be described thanks these conditions.

The bounds for $d(a, S)$ are obtained from the restriction of f to the linear manifold $a + N_2$, where N_2 is the orthogonal complement of the nullspace N_1 of $Df(a)$. Since $N_2 = \text{im}(Df(a)^*)$, we find for that restriction $f|_{N_2}$:

$$x \mapsto f(a + x), \quad x \in N_2 = \{Df(a)^* y \mid y \in Y\}.$$

If $f(a + Df(a)^* y) = 0$ then

$$d(a, S) \leq \|Df(a)^* y\|$$

In order to prepare and structure the proof of theorem 2.1 we start to formulate some lemmata.

LEMMA 2.1. Let N_1 be a closed subspace of Hilbertspace X and $N_2 = N_1^\perp$. Let α, β be positive reals and $w = w_1 + w_2 \in B(0, \beta)$ with $w_i \in N_i$ ($i = 1, 2$)

and $w_1 \neq 0$. Let be

$$K := \{(1 - \|w_1\|^{-1}\tau)w_1 + x_2 \in X \mid \tau \in [0, \|w_1\|], \\ x_2 \in N_2 \cap B(w_2, \alpha\tau)\} \quad (2.2)$$

a subset of a convex cone in X with top w and let

$$W := \{x_1 + x_2 \in X \mid \alpha\|x_1\| + \|x_2\| < \beta, x_i \in N_i, i = 1, 2\} \quad (2.3)$$

Then $K \subset B(0, \beta)$ if and only if $w \in W$.

PROOF. (i) Assume $w \in W$. If $x = x_1 + x_2 \in K$, then there exists a $\tau \in [0, \|w_1\|]$ such that $x_1 = (1 - \|w_1\|^{-1}\tau)w_1$ and $x_2 = w_2 + \varepsilon\alpha\tau v$ with $\varepsilon \in [0, 1)$ and v a unit vector in N_2 .

Hence we investigate

$$\|x\|^2 = \|x_1\|^2 + \|x_2\|^2 = (\|w_1\| - \tau)^2 + \|w_2 + \varepsilon\alpha\tau v\|^2 \\ \leq (\|w_1\| - \tau)^2 + (\|w_2\| + \alpha\tau)^2 =: g(\tau), \quad 0 \leq \tau \leq \|w_1\|.$$

Now on the interval $[0, \|w_1\|]$ the quadratic function g attains its maximum in one of the endpoints $0, \|w_1\|$. For $\tau = 0$ we obtain $g(0) = \|w_1\|^2 + \|w_2\|^2 < \beta^2$ since $w \in B(0, \beta)$ as given. For $\tau = \|w_1\|$ we obtain $g(\|w_1\|) = (\alpha\|w_1\| + \|w_2\|)^2 < \beta^2$ since $w \in W$ as assumed. So $x \in B(0, \beta)$ since $0 < g(\tau) < \beta^2$ for each $\tau \in [0, \|w_1\|]$.

(ii) Conversely, suppose $w \notin W$, i.e. $\alpha\|w_1\| + \|w_2\| \geq \beta$. Now consider $x := (1 - \|w_1\|^{-1}t)w_1 + x_2 \in K$ with $t = \|w_1\|$ and

$$x_2 = \begin{cases} \alpha\|w_1\|v, \text{ where } v \in M_2, \|v\| = 1, \text{ if } w_2 = 0 \\ w_2 + \alpha \frac{\|w_1\|}{\|w_2\|} w_2, \text{ if } w_2 \neq 0. \end{cases}$$

Then $x = x_2$ and thus $\|x\| = \alpha\|w_1\| + \|w_2\| \geq \beta$. So $x \notin B(0, \beta)$. □

The spherical symmetric forms in the next lemma make its proof to be a simple exercise.

LEMMA 2.2. Let $w \in X$ and W as in lemma 2.1. The distance $d(0, W^C)$ equals $\beta(1 + \alpha^2)^{\frac{1}{2}}$. □

In the proof of the main theorem 2.1 frequently the decomposition $X = N_1 \oplus N_2$, with $N_2 = N_1^\perp$, is used. Accordingly vectors x, y are decomposed: $x = x_1 + x_2, y = y_1 + y_2$ with $x_i, y_i \in N_i, i = 1, 2$. As appears, the subspace N_1 is the nullspace of the Frechet derivative at a given point a in the domain of a differentiable mapping.

THEOREM 2.1. Let X, Y be Hilbertspaces and $f: X \supset B(a, r) \rightarrow Y$ ($r > 0$) a Frechet differentiable mapping such that

- (i) $A := Df(a) \in L(X, Y)$ is surjective and $\|A^+\| \leq \lambda^{-1}$, where $A^+ \in L(X, Y)$ is the right inverse of A ;
- (ii) $\|Df(x) - Df(y)\| \leq L\|x - y\|, x, y \in B(a, r)$;
- (iii) $\|A^+ f(a)\| = \tilde{\gamma} \leq \gamma$;
- (iv) $\kappa := L\gamma \lambda^{-1} < \frac{1}{2}$ and $\rho := \lambda(1 - \sqrt{1-2\kappa})/L < r$.

Then the equation $f(x) = 0$ has a solution z in $B(a, \rho) \cap N_2$ which is unique in $B(a, \rho_2) \cap N_2$ where N_2 is the orthogonal complement of $N_1 := \ker(A)$ and

$$\rho_2 = \lambda(1 + \sqrt{1-2\kappa})/L. \tag{2.4}$$

The distance $d(a, S)$, where $S = \{x \in B(a, r) \mid f(x) = 0\}$, satisfies the inequalities

$$\rho_1 := \rho_3(1 + L^2 \rho_3^2 (\lambda - L\rho_3)^{-2})^{-\frac{1}{2}} < d(a, S) < \rho, \tag{2.5}$$

where

$$\rho_3 = \lambda(1 + \sqrt{1-2L\tilde{\gamma}\lambda^{-1}})/L \quad (2.6)$$

PROOF. Without loss of generality we may assume $a > 0$. The surjectivity of A implies that the restriction $A|_{N_2}$ of A to the closed subspace N_2 is bijective; by the Banach open mapping theorem its inverse is also continuous and equals the rightinverse A^+ of A : $AA^+y = y$ for each $y \in Y$. The rightinverse of A equals $A(AA^*)^{-1}$ [4]. Let $\varphi := f|_{N_2}$. This mapping satisfies the conditions of the Kantorovich theorem:

- (i) $D\varphi(0) = A|_{N_2} \in L(N_2, Y)$ is invertible and $\|D\varphi(0)^{-1}\| = \|A^+\| \leq \lambda^{-1}$;
- (ii) $\|D\varphi(x) - D\varphi(y)\| \leq \|Df(x) - Df(y)\| \leq L\|x - y\|$, $x, y \in N_2 \cap B(0, r)$;
- (iii) $\|D\varphi(0)^{-1} \varphi(0)\| = \tilde{\gamma} \leq \gamma$;
- (iv) $\kappa = L\gamma\lambda^{-1} < \frac{1}{2}$ and $\rho = \lambda(1 - \sqrt{1-2\kappa})/L < r$.

Hence the equation $\varphi(x) = 0$ has a solution z in $B(0, \rho) \cap N_2$ and this solution is unique in $B(0, \rho_2) \cap N_2$ with ρ_2 as given in (2.4) [8]. This z , zero of f in $B(0, \rho)$ is used to obtain a lowerbound ρ_1 of $d(0, S)$. Since Df is Lipschitz continuous on $B(0, r)$ we have [17]

$$\|f(z) - f(0) - Az\| \leq \frac{1}{2} L\|z\|^2$$

and consequently

$$\tilde{\gamma} = \|A^+f(0)\| = \|z + A^+(f(z) - f(0) - Az)\| \leq \|z\| + \frac{1}{2} L\lambda^{-1}\|z\|^2.$$

Hence the positive zero ρ_3 of the quadratic function

$$t \mapsto \frac{1}{2} L\lambda^{-1} t^2 + t - \tilde{\gamma}$$

is majorized by $\|z\|$. With (iv) we find $\|z\| < \rho < 2\gamma$ and consequently

$$L\|z\| < 2L\gamma < \lambda.$$

Let be $V := \{x \in X \mid \|x\| < \|z\|\}$ and

$$P(x) := Df(x)|_{N_1}, \quad Q(x) := Df(x)|_{N_2}, \quad x \in V.$$

If $y = y_1 + y_2$, $y_i \in N_i$ ($i = 1, 2$) then

$$Df(x)y = Df(x)(y_1 + y_2) = P(x)y_1 + Q(x)y_2.$$

Since $P(0) = 0$ we derive from the Lipschitz continuity

$$\|P(x)\| = \|P(x) - P(0)\| \leq \|Df(x) - Df(0)\| \leq L\|x\| \leq L\|z\|, \quad x \in V$$

and similarly

$$\|Q(x) - Q(0)\| \leq \|Df(x) - Df(0)\| \leq L\|z\|, \quad x \in V.$$

But $Q(0) = D\varphi(0)$ and so

$$\|(Q(x) - Q(0))Q(0)^{-1}\| \leq L\|z\| \lambda < 1, \quad x \in V,$$

which implies that $Q(x) = (I + (Q(x) - Q(0))Q(0)^{-1})Q(0)$ is invertible for each $x \in V$. With Banach's theorem [13] we obtain readily an upperbound, say α , of $\|Q(x)^{-1}P(x)\|$:

$$\|Q(x)^{-1}P(x)\| \leq \|Q(0)^{-1}\| \|P(x)\| / (1 - L\|z\|\lambda^{-1}) \leq \frac{L\|z\|}{\lambda - L\|z\|} =: \alpha, \quad x \in V \quad (2.7)$$

Let be $\Omega := \{x \in V \mid f(x) = 0\}$. From the uniqueness of the zero z of φ in $B(0, \rho) \cap N_2$ we conclude that $\|x\| \leq \|z\| \wedge x \in N_2 \wedge f(x) = 0$ implies $x = z$.

Now we assume $w = w_1 + w_2 \in \Omega$, with $w_i \in N_i$ ($i = 1, 2$) and $w_1 \neq 0$.

In our investigations we use the line ℓ in N_1 :

$$\ell := \{(1 - t\|w_1\|^{-1})w_1 \mid t \in \mathbb{R}\}$$

and the function F defined as follows:

$$\mathbb{R} \times N_2 \ni (t, y_2) \mapsto F(t, y_2) := f((1-t\|w_1\|^{-1})w_1 + y_2), t^2 + \|y_2\|^2 < r^2.$$

It is clear that $F(0, w_2) = 0$. Since

$$F(0, w_2 + h_2) - F(0, w_2) - Df(w)h_2 = o(\|h_2\|), (h_2 \rightarrow 0),$$

the derivative $D_2 F(0, w_2)$ of F in $(0, w_2)$ with respect to w_2 equals $Df(w)|_{N_2} = Q(w)$ and this derivative is a regular element of $\mathcal{L}(N_2, Y)$. In conformity with the implicit function theorem [20], there exists an interval $I \subset \mathbb{R}$ around 0 and a differentiable mapping $\psi: I \rightarrow N_2$ such that $\psi(0) = w_2$, $F(t, \psi(t)) = 0$ and too for each $t \in I$

$$D\psi(t) = -(D_2 F(t, \psi(t)))^{-1} D_1 F(t, \psi(t)).$$

(, $D_1 F(t, \psi(t))$ being the derivative of F in $(t, \psi(t))$ with respect to t). For δ small enough the differentiable curve

$$\{(1 - t\|w_1\|^{-1})w_1 + \psi(t) \in X \mid t \in (-\delta, \delta) \subset I\}$$

is a subset of V . On this interval $(-\delta, \delta)$ we have

$$\psi(t) - \psi(0) = - \int_0^t D_2 F(\tau, \psi(\tau))^{-1} D_1 F(\tau, \psi(\tau)) d\tau. \quad (2.8)$$

By the definition of F

$$D_1 F(t, y_2) = -P((1 - t\|w_1\|^{-1})w_1 + y_2)\|w_1\|^{-1} w_1$$

so

$$\|D_1 F(t, y_2)\| \leq \|P((1 - t\|w_1\|^{-1})w_1 + y_2)\|.$$

Hence for $|t| < \delta$ we obtain with (2.7)

$$\|D_2 F(t, \psi(t))^{-1} D_1 F(t, \psi(t))\| \leq \|Q((1-t\|w_1\|^{-1})w_1 + \psi(t))^{-1} P((1-t\|w_1\|^{-1})w_1 + \psi(t))\| \leq \alpha \quad (2.9)$$

With this upperbound we find easily from (2.8)

$$\|\psi(t) - w_2\| \leq \alpha|t|, \quad |t| < \delta. \quad (2.10)$$

Let us suppose that (, compare with (2.3),)

$$w \in W := \{x = x_1 + x_2 \in V \mid \alpha\|x_1\| + \|x_2\| < \|z\|, x_i \in N_i, i = 1, 2\} \quad (2.11)$$

Then, as follows from lemma 2.1, for each $t \in [0, \|w_1\|]$

$$(1 - t\|w_1\|^{-1})w_1 + x_2 \in B(0, \|z\|) \quad (2.12)$$

if $x_2 \in B(w_2, \alpha t)$. So if $w \in W$ than the function ψ (, solving y_2 from the equation $F(t, y_2) = 0$ as a function of t ,) can be continued to the right until $t = \|w_1\|$. Analogously to (2.10) one finds for this extended mapping ψ :

$$\|\psi(\|w_1\|) - w_2\| \leq \alpha\|w_1\|.$$

But

$$0 = F(\|w_1\|, \psi(\|w_1\|)) = f(\psi(\|w_1\|)) = \varphi(\psi(\|w_1\|)),$$

for $\psi(\|w_1\|) \in N_2$. With (2.12) we obtain

$$\psi(\|w_1\|) \in B(0, \|z\|).$$

This conclusion contradicts that z is the unique zero of ψ in the ball $B(0, \rho_2) \cap N_2$. The contradiction proves the incorrectness of assumption (2.11)! We have proved that

$$\Omega \subset W^c = \{x = x_1 + x_2 \in X \mid \alpha\|x_1\| + \|x_2\| \geq \|z\|, x_i \in N_i, i = 1, 2\} \quad (2.13)$$

The distance $d(0, \Omega)$ is minorized by $d(0, W^C)$; with lemma 2.2 one finds

$$d(0, \Omega) \geq \frac{\|z\|}{\sqrt{1+\alpha^2}} = \frac{\|z\|(\lambda - L\|z\|)}{\sqrt{L^2\|z\|^2 + (\lambda - L\|z\|)^2}} = \|z\| \left(1 + \left(\frac{v}{1-v}\right)^2\right)^{-\frac{1}{2}},$$

where $v = L\|z\|/\lambda \in [L\rho_3\lambda^{-1}, L\rho\lambda^{-1}) \subset (0, 1)$. Since the function

$$v \rightarrow \left(1 + \left(\frac{v}{1-v}\right)^2\right)^{-\frac{1}{2}}$$

on this interval attains its minimum in the endpoint $L\rho_3\lambda^{-1}$, we finally obtain

$$d(0, \Omega) \geq \rho_3(1 + L^2\rho_3^2(\lambda - L\rho_3)^{-2})^{-\frac{1}{2}} = \rho_1. \quad \square$$

REMARK 1. Under the same conditions as in theorem 2.1, if $\kappa < \frac{1}{2}$, then the iterates

$$x_{k+1} := x_k - A^*(AA^*)^{-1} f(x_k), \quad k \in \mathbb{N}, x_1 := 0, \quad (2.14)$$

are defined for all k , and converge to $z \in B(0, \rho) \cap N_2$ [20].

This modified Newton's method gives a linearly convergent sequence $\{x_k\} \subset B(0, \rho_2) \cap N_2$ and

$$\|x_k - z\| \leq 2 \frac{\lambda}{L} (1 - \sqrt{1-2\kappa})^{k+1}. \quad \square$$

REMARK 2. The bounds for the zero of $f|_{N_2}$ can be obtained by a direct analysis of the mapping $\varphi = f|_{N_2}$. If $x \in N_2$, then there exists a unique $y \in Y$ such that $x = A^*y$; hence if $\hat{y} \in Y$ is a zero of $K: Y \rightarrow Y$, where $K(y) = f(A^*y)$, then $z = A^*\hat{y} \in N_2$ is a zero of f . That \hat{y} can be obtained with modified Newton's method:

$$y_{k+1} := y_k - (AA^*)^{-1} f(A^*y_k), \quad k \in \mathbb{N}, y_1 := 0 \quad (2.15)$$

Evidently, the sequences $\{x_k\}$ and $\{y_k\}$, given in (2.14) and (2.15) are

strongly related: $x_k = A^* y_k$.

□

REMARK 3. Quite the same results can be derive for $f: X \rightarrow Y$ where X and Y are Hilbert space over the field of complex numbers. This will occur in section 4 where the results of this section are applied to the eigenproblem. □

3. DISTRIBUTION OF ERRORS IN A LINEAR CONTINUOUS SYSTEM

3.1.

Let be given a real continuous function $x \in C(I)$, I being the interval $[0,1]$ and a real differentiable function $y \in C^1(I)$. Further we assume to have at our disposal the numbers $a, b \in \mathbb{R}$. This quadruple (a, b, x, y) defines residual $r \in C(I)$:

$$r(t) := \dot{y}(t) - ay(t) - bx(t), \quad 0 \leq t \leq 1. \quad (3.1)$$

In order to evaluate how far the pair (x, y) fits in the scalar linear model

$$\dot{y} = ay + bx \quad (3.2)$$

we introduce perturbations

$$\varphi = (\alpha, \beta, \rho, \xi, \eta) \in X := C(I)^4 \times C^1(I) \quad (3.3)$$

such that

$$\dot{y}(t) + \dot{\eta}(t) = (a+\alpha(t))(y(t)+\eta(t)) + (b+\beta(t))(x(t)+\xi(t)) + \rho(t) \quad (3.4)$$

So this quintuple φ has to satisfy the equation $f(\varphi) = 0$ where $f: X \rightarrow Y$, $Y := C(I)$, with

$$\begin{aligned} f(\varphi)(t) &= f(\alpha, \beta, \rho, \xi, \eta)(t) = \dot{y}(t) + \dot{\eta}(t) - (a+\alpha(t))(y(t)+\eta(t)) - \\ &\quad (b+\beta(t))(x(t)+\xi(t)) - \rho(t) \\ &= r(t) + \dot{\eta}(t) - a\eta(t) - b\xi(t) - \alpha(t)y(t) - \beta(t)x(t) - \\ &\quad \rho(t) - \alpha(t)\eta(t) - \beta(t)\xi(t). \end{aligned} \quad (3.5)$$

The modification of (3.2) into

$$\dot{\eta}(t) = (a+\alpha(t))\eta(t) + \beta(t)(x(t)+\xi(t)) + b\xi(t) + \alpha y(t) + \rho(t) - r(t) \quad (3.6)$$

gives a weakly nonlinear system [6].

The functions α and β are considered to be fluctuations of the structural parameters a and b respectively. The functions ξ and η are perturbations of the given functions x and y and ρ is a corrected residual. In this view the description of the error in the model with residual r , has been replaced by a description with the fluctuations α and β , the perturbations ξ and η and the corrected residual ρ . In a natural way the question rises to determine the minimal $\varphi = (\alpha, \beta, \rho, \xi, \eta)$, with respect to some norm, such that a , b , x and y fit in system (3.4).

The same problem can be seen as an example of optimal control. The determination of an minimal $\varphi = (\alpha, \beta, \rho, \xi, \eta)$ (, with respect to some norm) subject to condition (3.6) is just a weakly nonlinear tracking problem. In the linear state equation (3.6) the state η and the controls α , β , ξ and ρ occur in a bilinear way [1,6,11,12,21].

3.2.

In this section we apply the approach of the preceding section for the estimation of that minimal φ .

As we see from (3.6)

$$f(\varphi_1 + \varphi_2)(t) = f(\varphi_1)(t) + Df(\varphi_1) \varphi_2(t) + R(\varphi_1; \varphi_2)(t) \quad (3.7)$$

where

$$Df(\varphi_1)\varphi_2(t) = \dot{\eta}_2 - (a+\alpha_1)\eta_2 - (b+\xi_1)\xi_2 - \alpha_2(y+\eta_1) - \beta_2(x+\xi_1) - \rho_2 \quad (3.8)$$

and

$$R(\varphi_1; \varphi_2)(t) = -\alpha_2\eta_2 - \beta_2\xi_2 \quad (3.9)$$

with $\varphi_i = (\alpha_i, \beta_i, \rho_i, \xi_i, \eta_i)$, $i = 1, 2$.

Both φ_1 and increment φ_2 are elements of the productspace X with inner-product

$$\langle \varphi_1, \varphi_2 \rangle := (\alpha_1, \alpha_2) + (\beta_1, \beta_2) + (\rho_1, \rho_2) + (\xi_1, \xi_2) + (\eta_1, \eta_2),$$

$$\varphi_1, \varphi_2 \in X,$$

where $\varphi_i = (\alpha_i, \beta_i, \rho_i, \xi_i, \eta_i)$, $i = 1, 2$.

With the L_2 -norms in the prehilbert spaces X and Y we find for $f: X \rightarrow Y$ that

$$R(\varphi_1, \varphi_2) \|\varphi_2\|^{-1} \rightarrow 0 \quad (\varphi_2 \rightarrow 0),$$

for each $\varphi_1 \in X$. Nevertheless f is not Frechet differentiable at $\varphi_1 \in X$, for the linear mapping $Df(\varphi_1): X \rightarrow Y$ is not bounded; on the other hand f is Gateaux differentiable at each $\varphi \in X$.

The constant term r of $f(\varphi)$ equals $f(0)$ and the linear part, relatively the L_2 -norms in X and Y , is $A\varphi(t) := Df(0)\varphi(t) = \dot{\eta}(t) - a\eta(t) - b\xi(t) - \alpha(t)y(t) - \beta(t)x(t) - \rho(t)$.

The domain X of the Gateaux differential A of f at 0 is dense in $H := L_2(I)$ and its range Y is dense in $L_2(I)$.

Indeed the mapping f is surjective as is A : for each $p \in Y$ we have $f(0, 0, r-p, 0, 0) = A(0, 0, -p, 0, 0) = p$.

Now we start to construct the adjoint A^* of A . Suppose $\psi \in D(A^*)$ and let be $A^*\psi = (\alpha_1, \beta_1, \rho_1, \xi_1, \eta_1)$. This means that for each $(\alpha, \beta, \rho, \xi, \eta) \in X$ holds

$$\begin{aligned} (A(\alpha, \beta, \rho, \xi, \eta), \psi) &= \int_0^1 (\dot{\eta} - a\eta - b\xi - \alpha y - \beta x - \rho)(t) \psi(t) dt \\ &= ((\alpha, \beta, \rho, \xi, \eta), (\alpha_1, \beta_1, \rho_1, \xi_1, \eta_1)) = \int_0^1 (\alpha\alpha_1 + \beta\beta_1 + \rho\rho_1 + \xi\xi_1 + \eta\eta_1)(t) dt \end{aligned}$$

Thus for each $(\alpha, \beta, \rho, \xi, \eta) \in X$ we have

$$\int_0^1 [(\dot{\eta} - a\eta)\psi - \eta\dot{\eta}_1](t) dt = \int_0^1 [(b\psi + \xi_1)\xi + (y\psi + \alpha_1)\alpha + (x\psi + \beta_1)\beta + (\psi + \rho_1)\rho](t) dt \quad (3.10)$$

If in the quintuple $(\alpha, \beta, \rho, \xi, \eta) \in X$ each element is taken zero except respectively α , β , ρ and ξ we find

$$\alpha_1(t) = -y(t)\psi(t), \beta_1(t) = -x(t)\psi(t), \rho_1(t) = -\psi(t), \xi_1(t) = -b\psi(t),$$

$$0 \leq t \leq 1 \quad (3.11)$$

If $\alpha = \beta = \rho = \xi = 0$ we find with (3.10) that for each $\eta \in C^1(I)$:

$$\int_0^1 [(\dot{\eta} - a\eta)\psi - \eta\eta_1] dt = 0.$$

Since the adjoint of $\eta \mapsto \dot{\eta}$ is the mapping $\psi \mapsto -\dot{\psi}$ with $\psi(0) = \psi(1) = 0$ [10] and that of $\eta \mapsto a\eta$ is $\psi \mapsto a\psi$ we find

$$\eta_1(t) = -\dot{\psi}(t) - a\psi(t), \psi(0) = \psi(1) = 0.$$

So

$$\left\{ \begin{array}{l} D(A^*) = \{\psi \in C^1(I) \mid \psi(0) = \psi(1) = 0\} \\ A^*\psi(t) = (-y(t)\psi(t), -x(t)\psi(t), -\psi(t), -b\psi(t), -\dot{\psi}(t) - a\psi(t)), 0 \leq t \leq 1. \end{array} \right.$$

Thus far the mappings f and $Df(\psi)$ have been considered on the domain X ; the nonlinear term $\beta\xi$ in the expression (3.5) of $f(\varphi)$ prevents f to be defined on the Hilbert space $L_2(1)^5 \times H^1(I)$. Consistently, we now consider the restriction of f to

$$N_2 := R(A^*) = \{(-y\psi, -x\psi, -\psi, -b\psi, -\dot{\psi} - a\psi) \mid \psi \in C^1(I), \psi(0) = \psi(1) = 0\}.$$

A usable upperbound of $\inf\{\|\varphi\|_2 \mid \varphi \in X, f(\varphi) = 0\}$ can be derived from a zero of $f|_{N_2}$.

Let $\varphi \in N_2$. Then there exists a $\psi \in C^1(I)$, with $\psi(0) = \psi(1) = 0$, such that

$$\varphi(t) = (-y(t)\psi(t), -x(t)\psi(t), -\psi(t), -b\psi(t), -\dot{\psi}(t) - a\psi(t)), t \in I.$$

(3.12)

Hence $f(\emptyset) = 0$ implies

$$\left\{ \begin{array}{l} K(\psi)(t) := r(t) - \dot{\psi}(t) + g(t)\psi(t) - y(t)\psi(t)\dot{\psi}(t) - k(t)\psi^2(t) = 0 \end{array} \right. \quad (3.13)$$

$$\left\{ \begin{array}{l} \psi(0) = \psi(1) = 0 \end{array} \right. \quad (3.14)$$

where

$$g(t) := a^2 + b^2 + 1 + x^2(t) + y^2(t), \quad k(t) := ay(t) + bx(t), \quad t \in I.$$

The solution of the equation $f(\psi) = 0$ in N_2 can be derived with (3.12) from the solution of the two point boundary value problem (TPBVP) (3.13), (3.14). So we investigate the existence problem for this TPBVP. Instead of the Newton approach we make use of the Banach fixed point theorem in order to proof the existence and to estimate the solution of this TPBVP [3], [5].

3.3.

For the proof of the existence of a solution $\psi \in C^1([0,1])$ of TPBVP (3.13), (3.14) we consider the differential operator $P:V \rightarrow C([0,1])$, where

$$\left\{ \begin{array}{l} V := \{ \psi \in C^2[0,1] \mid \psi(0) = \psi(1) = 0 \} \end{array} \right. \quad (3.15)$$

$$\left\{ \begin{array}{l} P\psi = w \Leftrightarrow \dot{\psi}(t) - g(t)\psi(t) = w(t), \quad t \in [0,1] \end{array} \right. \quad (3.16)$$

Since $g(t) > 0$, $0 \leq t \leq 1$, the linear operator P is regular [7].

With respect to the supremum norm $\|\cdot\|_\infty$ on $C[0,1]$ operator P has a bounded inverse $P^{-1}: C[0,1] \rightarrow V$ such that [10]

$$\|P^{-1}\|_\infty \leq \frac{1}{m(g)}, \quad (3.17)$$

where

$$m(g) := \inf\{g(t) \mid t \in [0,1]\}. \quad (3.18)$$

Thus for the solution u of TPBVP

$$\bar{\psi} - g\psi = w, \quad \psi(0) = \psi(1) = 0$$

we have

$$\|\psi\|_{\infty} \leq \frac{\|w\|_{\infty}}{m(g)}$$

Further holds [7]

$$\psi(t) = \int_0^1 G(t,s) (g(s)\psi(s) + w(s)) ds, \quad (3.19)$$

where

$$G(t,s) = \begin{cases} (t-1)s, & 0 \leq s \leq t \leq 1 \\ t(s-1) & 0 \leq t \leq s \leq 1 \end{cases} \quad (3.20)$$

Now we investigate the mapping $T: C^1[0,1] \rightarrow C^1[0,1]$, where

$$Tv = u \Leftrightarrow \begin{cases} \bar{u} - gu = r - yv\hat{v} - kv^2 \\ u(0) = u(1) = 0. \end{cases} \quad (3.21)$$

Hence

$$u = Tv = P^{-1}(r - yv\hat{v} - kv^2) \quad (3.22)$$

and

$$u(t) = \int_0^1 G(t,s) (gu + r - yv\hat{v} - kv^2)(s) ds \quad (3.23)$$

Consequently any point $\psi \in C^1[0,1]$ is a solution of TPBVP (3.13), (3.14) if and only if $\psi = T\psi$.

With respect to the norm

$$\|v\|_S := \|v\|_{\infty} + \|\hat{v}\|_{\infty}, \quad v \in C^1[0,1] \quad (3.24)$$

the linear space $C^1[0,1]$ is complete.

In $C^1[0,1], \|\cdot\|_s$ we determine a closed ball $W := \overline{B(0,p)}$ such that $T(W) \subset W$. Equation (3.22) implies

$$\|u\|_\infty \leq (\|r\|_\infty + \|y\|_\infty \|v\|_\infty \|\dot{v}\|_\infty + \|k\|_\infty \|v\|_\infty^2) / m(g). \quad (3.25)$$

From (3.19) we derive

$$\dot{u}(t) = \int_0^1 \frac{\partial G}{\partial t}(t,s) (gu + r - yv\dot{v} - kv^2)(s) ds,$$

where

$$\frac{\partial G}{\partial t}(t,s) = \begin{cases} s, & 0 \leq s \leq t \leq 1 \\ s-1, & 0 \leq t \leq s \leq 1. \end{cases}$$

Consequently

$$\|\dot{u}\|_\infty \leq \frac{3}{2} [\|g\|_\infty \|u\|_\infty + \|r\|_\infty + \|y\|_\infty \|v\|_\infty \|\dot{v}\|_\infty + \|k\|_\infty \|v\|_\infty^2]. \quad (3.26)$$

Addition of (3.25) and (3.26) gives

$$\begin{aligned} \|u\|_s &\leq c(g) [\|r\|_\infty + \|y\|_\infty \|v\|_\infty \|\dot{v}\|_\infty + \|k\|_\infty \|v\|_\infty^2] \\ &\leq c(g) [\|r\|_\infty + (\|y\|_\infty + \|k\|_\infty) \|v\|_s^2] \end{aligned} \quad (3.27)$$

where

$$c(g) := \frac{3}{2} \left(\frac{\|g\|_\infty}{m(g)} + 1 \right) + \frac{1}{m(g)}. \quad (3.28)$$

Thus $\|v\|_s \leq p$ implies $\|Tv\|_s \leq p$ if

$$c(g) (\|r\|_\infty + (\|y\|_\infty + \|k\|_\infty) p^2) < p.$$

The last inequality is satisfied if and only if

$$p_1 \leq p \leq p_2 \quad (3.29)$$

where

$$p_i = \frac{1 + (-1)^i \sqrt{1 - 4c^2(g)(\|y\|_\infty + \|k\|_\infty)\|r\|_\infty}}{2c(g)(\|y\|_\infty + \|k\|_\infty)}, \quad i = 1, 2. \quad (3.30)$$

Evidently, residual r has to be sufficient small:

$$\|r\|_\infty \leq (4c^2(g)(\|y\|_\infty + \|k\|_\infty))^{-1} \quad (3.31)$$

Now we investigate $T:W \rightarrow W$ ($p_1 \leq p \leq p_2$) on its contraction property.

Let be $u_i = Tv_i$, $v_i \in W$, $i = 1, 2$.

Then

$$u_1 - u_2 = P^{-1}(y(v_2 \hat{v}_2 - v_1 \hat{v}_1) + k(v_2^2 - v_1^2))$$

and

$$\dot{u}_1(t) - \dot{u}_2(t) = \int_0^1 \frac{\partial G}{\partial t}(t, s)(g(u_1 - u_2) + y(v_2 \hat{v}_2 - v_1 \hat{v}_1) + k(v_2^2 - v_1^2))(s) ds.$$

Since

$$\begin{cases} v_2 \hat{v}_2 - v_1 \hat{v}_1 = v_2(\hat{v}_2 - \hat{v}_1) + (v_2 - v_1)\hat{v}_1 \\ v_2^2 - v_1^2 = (v_2 + v_1)(v_2 - v_1), \end{cases}$$

evidently

$$\begin{cases} \|v_2 \hat{v}_2 - v_1 \hat{v}_1\|_\infty \leq (\|v_1\|_s + \|v_2\|_s)\|v_1 - v_2\|_s \\ \|v_2^2 - v_1^2\|_\infty \leq (\|v_1\|_s + \|v_2\|_s)\|v_1 - v_2\|_s. \end{cases}$$

Then

$$\|Tv_1 - Tv_2\|_\infty = \|u_1 - u_2\|_\infty \leq \frac{1}{m(g)} (\|y\|_\infty + \|k\|_\infty) (\|v_1\|_s + \|v_2\|_s) \|v_1 - v_2\|_s \quad (3.32)$$

and

$$\begin{aligned} |\dot{u}_1(t) - \dot{u}_2(t)| &\leq \frac{3}{2} (\|g\|_\infty \|u_1 - u_2\|_\infty + \|y\|_\infty \|v_2 - v_1\|_\infty + \|k\|_\infty \|v_2^2 - v_1^2\|_\infty) \\ &\leq \frac{3}{2} \left(\frac{\|g\|_\infty}{m(g)} + 1 \right) (\|y\|_\infty + \|k\|_\infty) (\|v_1\|_S + \|v_2\|_S) \|v_1 - v_2\|_S. \end{aligned} \quad (3.33)$$

Addition of (3.32) and (3.33) gives,

$$\begin{aligned} \|Tv_1 - Tv_2\|_S &\leq c(g) (\|y\|_\infty + \|k\|_\infty) (\|v_1\|_S + \|v_2\|_S) \|v_1 - v_2\|_S \\ &\leq 2pc(g) (\|y\|_\infty + \|k\|_\infty) \|v_1 - v_2\|_S. \end{aligned}$$

So the mapping $T:W \rightarrow W$ is a contraction if

$$p < p_3 := \frac{1}{2c(g) (\|y\|_\infty + \|k\|_\infty)} \quad (3.33)$$

Remark that, provided condition (3.31) is satisfied, $p_1 \leq p_3 \leq p_2$. Now we have reached the following result.

The mapping $T:W = \overline{B(0,p)} \rightarrow W$ is a contraction for each $p \in [p_1, p_3]$ if

$$\|r\|_\infty \leq \frac{1}{4c^2(g) (\|y\|_\infty + \|k\|_\infty)}.$$

Under these conditions T has a unique fixed point ψ in $\overline{B(0,p)}$. Moreover, the sequence $\{u_n\}$, with $u_0 = 0$ and

$$u_n = Tu_{n-1}, \quad n \in \mathbb{N},$$

converges in $C^1[0,1], \|\cdot\|_S$ to that solution ψ of the equation $u = Tu$. For that solution ψ holds the inequality

$$\|\psi\|_S \leq \frac{\|u_1\|_S}{1 - 2pc(g) (\|y\|_\infty + \|k\|_\infty)}$$

The first iterate u_1 is solution of the TPBVP

$$\ddot{u}_1 - gu_1 = r, \quad u_1(0) = u_1(1) = 0.$$

As a consequence of (3.27)

$$\|u_1\|_\infty \leq c(g)\|r\|_\infty.$$

Therefore

$$\|\psi\|_s \leq \frac{c(g)}{1-2pc(g)(\|y\|_\infty + \|k\|_\infty)} \|r\|_\infty. \quad (3.34)$$

With $p = p_1$ finally we have

$$\|\psi\|_s \leq \frac{c(g)}{1-2p_1c(g)(\|y\|_\infty + \|k\|_\infty)} \|r\|_\infty.$$

For the solution $\phi := (-y\psi, -x\psi, -\psi, -b\psi, -\dot{\psi}-a\psi)$ of equation $f(\phi) = 0$ we have

$$\begin{aligned} \int_0^1 \phi^2(t) dt &= \int_0^1 [(y^2+x^2+1+b^2+a^2)\psi^2 + \dot{\psi}^2 + 2a\psi\dot{\psi}](t) dt \\ &= \int_0^1 (g(t)\psi^2(t) + \dot{\psi}^2(t)) dt \end{aligned}$$

since $\psi(0) = \psi(1) = 0$. Now

$$\|\phi\|_2^2 \leq \|g\|_\infty (\|\psi\| + \|\dot{\psi}\|)^2 = \|g\|_\infty \|\psi\|_s^2$$

Consequently

$$\|\hat{\phi}\|_s \leq \frac{c(g)\sqrt{\|g\|_\infty}}{1-2p_1c(g)(\|y\|_\infty + \|k\|_\infty)} \|r\|_\infty.$$

This result is summarized in

THEOREM 3.1. Let be given $a, b \in \mathbb{R}$, $x \in C(I)$ and $y \in C^1(I)$, where $I = [0,1]$.

Let

$$r(t) := \dot{y}(t) - ay(t) - bx(t), \quad g(t) := a^2 + b^2 + 1 + x^2(t) + y^2(t), \quad k(t) =$$

$$ay(t) + bx(t), \quad t \in I.$$

If

$$4\|r\|_{\infty} c^2(g) (\|y\|_{\infty} + \|k\|) < 1,$$

where

$$c(g) = \frac{3}{2} \left(\frac{\|g\|_{\infty}}{m(g)} + 1 \right) + \frac{1}{m(g)}$$

with

$$m(g) = \inf\{g(t) \mid t \in I\},$$

then there exists a perturbation $\phi = (\alpha, \beta, \rho, \xi, \eta) \in C(I)^4 \times C^1(I)$ such that

$$\dot{y} + \dot{\eta} = (a+\alpha)(y+\eta) + (b+\beta)(x+\xi) + \rho$$

and

$$\|\phi\|_2 \leq \frac{c(g) \sqrt{\|g\|_{\infty}}}{1 - 2p_1 c(g) (\|y\|_{\infty} + \|k\|_{\infty})} \|r\|_{\infty}$$

where

$$p_1 = \frac{1 - \sqrt{1 - 4c^2(g) (\|y\|_{\infty} + \|k\|_{\infty}) \|r\|_{\infty}}}{2c(g) (\|y\|_{\infty} + \|k\|_{\infty})} .$$

□

REMARK. In a next memorandum we will analyse the vectorial analogue of the distribution problem and also will be described a numerical method for the computation of these suboptimal fluctuations and perturbations.

4. A POSTERIORI ERROR BOUNDS FOR AN EIGENPAIR AND NUMERICAL RESULTS.

In this section we apply the results of section two to the eigenproblem. We assume to have at our disposal an approximate eigenproblem solution (λ, x) for a given complex $n \times n$ matrix. For instance, this eigenpair can be seen as the result of some eigenvalue computations performed with finite arithmetic.

Let be $r := Ax - \lambda x$, the residual vector corresponding to the approximate eigenpair (λ, x) ; further we assume x to be normalized in length: $x^* x = 1$. The pair (λ, x) has been afflicted with errors $(\epsilon, \epsilon \neq 0)$, so at the moment we introduce perturbations (δ, h) in order to compensate these errors, i.e. $(\lambda + \delta, x + h)$ has to be an exact eigenpair for matrix A .

For the analysis of the bounds for the necessary corrections (δ, h) we introduce the nonlinear, smooth function $f : C \times C^n \rightarrow C^n$, defined by

$$f(\delta, h) := A(x+h) - \lambda(x+h). \quad (4.1)$$

Then, $f(0,0) = r$ and for $P := Df(0,0)$ we find

$$P(\delta, h) = (A - \lambda I)h - \delta x. \quad (4.2)$$

With respect to the usual innerproducts in C^n and $C^{n+1} \sim C \times C^n$ we derive for the adjoint $P^* : C^n \rightarrow C \times C^n$ of Q :

$$\begin{aligned} (P^* k, (\delta, h)) &= (k, P(\delta, h)) = (k, (A - \lambda I)h - \delta x) \\ &= (k, -x)\bar{\delta} + ((A - \lambda I)^* k, h), \quad \delta \in C; h, k \in C^n. \end{aligned}$$

So $P^* k = ((k, -x), (A - \lambda I)^* k) \in C \times C^n$ and with (4.2) we obtain

$$PP^* k = (A - \lambda I)(A - \lambda I)^* k - (k, -x)x = (A - \lambda I)(A - \lambda I)^* k + xx^* k.$$

For the complex $n \times n$ matrix QQ^* we have found

$$PP^* = (A - \lambda I)(A - \lambda I)^* + xx^*. \quad (4.3)$$

If λ is not an eigenvalue of A , then PP^* is regular. If (λ, x) is an eigen-

pair, then the algebraic multiplicity of λ is decisive for the regularity of PP^* .

LEMMA 4.1. Let be (λ, x) an eigenpair of matrix A . Then $PP^* = (A-\lambda I)(A-\lambda I)^* + xx^*$ is regular if and only if the algebraic multiplicity of λ is equal to one.

PROOF. Without loss of generality we may assume $A-\lambda I$ to be an uppertriangular matrix and x to be e_1 , the first unit vector. Moreover, as concerns the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of A we assume $|\lambda_2 - \lambda| \leq |\lambda_3 - \lambda| \leq \dots \leq |\lambda_n - \lambda|$. Let

$$P = (A - \lambda I : x) = \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} & 1 \\ 0 & \lambda_2 - \lambda & a_{23} & \dots & a_{2n} & 0 \\ 0 & 0 & \lambda_3 - \lambda & \dots & a_{3n} & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_n - \lambda & 0 \end{pmatrix}.$$

Evidently, $\text{rank } P = n$ if and only if $\lambda_2 \neq \lambda$. □

After these preparations an obvious application of theorem 2.1 is formulated in

THEOREM 4.1. Let $f, P : C \times C^n \rightarrow C^n$ be defined by (4.1) and (4.2). If

- (i) $\text{rank } P = n, \| (PP^*)^{-1} \|_2^{1/2} = \mu_n^{-1} \leq \mu^{-1}$, μ_n being the smallest singular value of P ;
- (ii) $(r^*(PP^*)^{-1}r)^{1/2} = \tilde{\gamma} < \gamma$, where $r = f(0,0) = Ax - \lambda x$;
- (iii) $\kappa = \gamma \mu^{-1} < \frac{1}{2}$,

then the distance d of $(0,0)$ to $S = \{(\delta, h) \in C \times C^n | f(\delta, h) = 0\}$ satisfies the inequalities

$$\rho_1 := \rho_3 \left(1 + \frac{\rho_3^2}{(\mu - \rho_3)^2} \right)^{-\frac{1}{2}} < d < \rho,$$

where

$$\rho := \mu(1-\sqrt{1-2\kappa}), \rho_3 := \mu(\sqrt{1+2\tilde{\gamma}/\mu}-1).$$

PROOF. The only thing missing for the application of theorem 2.1 is the Lipschitz constant L of the Lipschitz continuous derivative

$$DF : C \times C^n \rightarrow M_{n,n+1}. \text{ Evidently } Df(\alpha, a)(\delta, h) = (A - (\lambda + \alpha)I)h - \delta(x + a).$$

Hence

$$\| (Df(\alpha_2, a_2) - Df(\alpha_1, a_1))(\delta, h) \|_2 = \| (\alpha_1 - \alpha_2)h + \delta(a_1 - a_2) \|_2.$$

Since

$$\begin{aligned} \| (\alpha_1 - \alpha_2)h + \delta(a_2 - a_1) \|_2^2 &= |\alpha_1 - \alpha_2|^2 \|h\|_2^2 + |\delta|^2 \|a_2 - a_1\|_2^2 + \operatorname{Re}\{(\alpha_1 - \alpha_2)\bar{\delta}(h, a_2 - a_1)\} \\ &\leq |\alpha_1 - \alpha_2|^2 \|h\|_2^2 + |\delta|^2 \|a_2 - a_1\|_2^2 + 2|\alpha_1 - \alpha_2| |\delta| \|a_2 - a_1\|_2^2 \\ &\leq |\alpha_1 - \alpha_2|^2 \|h\|_2^2 + |\delta|^2 \|a_2 - a_1\|_2^2 + |\alpha_1 - \alpha_2|^2 |\delta|^2 + \|a_2 - a_1\|_2^2 \|h\|_2^2 \\ &= (|\alpha_1 - \alpha_2|^2 + \|a_2 - a_1\|_2^2) (|\delta|^2 + \|h\|_2^2). \end{aligned}$$

With well chosen pair (δ, h) the equality signs occur, consequently

$$\| Df(\alpha_2, a_2) - Df(\alpha_1, a_1) \|_2 \leq (|\alpha_2 - \alpha_1|^2 + \|a_2 - a_1\|_2^2)^{\frac{1}{2}}.$$

This result implies that $L = 1$ is a Lipschitz constant for the derivative Df . Hence, with $L = 1$, theorem 2.1 immediately leads to this particular case. \square

So far only we dealt about an analytical expression for the error bounds.

But above all the Newton-Kantorovich approach provides a possibility to improve the approximate eigenvalue solution $(\lambda_0, x_0) := (\lambda, x)$.

Numerical results, obtained for F_{11} , the Frank matrix of order eleven, illustrate the performance of modified Newton's method in $(\ker P)^1$:

$$(\delta_k, h_k) = (\delta_{k-1}, h_{k-1}) - P^+ f(\delta_{k-1}, h_{k-1}), \quad k \in \mathbb{N}. \quad (4.4)$$

The general form of the Frank matrices is adequately illustrated by F_5

which is

$$\begin{pmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 4 & 3 & 2 & 1 \\ 0 & 3 & 3 & 2 & 1 \\ 0 & 0 & 2 & 2 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

"It can be shown that if λ_i is an eigenvalue then λ_i^{-1} is also an eigenvalue so that F_{2n+1} has one eigenvalue equal to unity. In general the larger eigenvalues are well-conditioned while the smaller ones are quite ill-conditioned." [16].

We used a single-precision approximation (λ_0, x_0) of the eigenpair for the smallest eigenvalue $\tilde{\lambda} = 0.034\ 625\ 161\ 711$ where $\lambda_0 = 0.034\ 623$. This (λ_0, x_0) was obtained by a NAG-implementation of the QR algorithm.

In our implementation of the Newton-algorithm (4.4) we used double length precision: long reals. With the usual inner product in C^n , the Newton correction $P^+ f(\delta_{k-1}, h_{k-1})$ was computed in the following way.

Firstly, the QR-decomposition is performed of the $(n+1) \times n$ matrix P^* with Householder reflections:

$$P^* = \begin{pmatrix} (A - \lambda I)^* \\ x^* \end{pmatrix} = QR.$$

Then

$$P^+ = P^* (P P^*)^{-1} = QR (R^* Q^* QR)^{-1} = Q (R^*)^{-1}.$$

Hence

$$(\delta_k, h_k) := (\delta_{k-1}, h_{k-1}) - Q (R^*)^{-1} f(\delta_{k-1}, f_{k-1}), \quad k \in \mathbb{N}. \quad (4.5)$$

So in each sweep the lower triangular system

$$R^* z_{k-1} = f(\delta_{k-1}, f_{k-1}), k \in \mathbb{N}$$

has to be solved.

In Table 1 we give the results for the sequence $\{\lambda_k\}$ where $\lambda_k = \lambda_0 + \delta_k$.

k	λ_k									
0	0.034	623								
1	0.034	625	161	711	439	764	197	778	019	752
2	0.034	625	161	711	425	640	061	346	944	761
3	0.034	625	161	711	425	620	857	605	864	163
4	0.034	625	161	711	425	620	855	414	507	969
5	0.034	625	161	711	425	620	855	414	291	505
6	0.034	625	161	711	425	620	855	414	291	351
correct numbers:	6				13	16	20	24		27

These numbers reflect the quadratic improvement in the first sweep of the modified method of Newton and they are in accordance with the figures of Symon and Wilkinson. [16] Evidently, if starting with the same (λ, x) and using single precision accuracy, one needs accumulation of innerproducts in double precision as an adequate precaution for otherwise the theoretic-al results will be invalidated by rounding errors.

Up to now, we have formulated error bounds in the forward sense. But theorem 2.1 also can be used to obtain error bounds in the mixed sense: both a perturbation H of the given matrix A and a perturbation h of the computed eigenvector x (, or of the computed eigenpair (λ, x) ,) are utilized in order to fulfill the eigenvalue-equation $(A+H)(x+h) = \lambda(x+h)$. The mixed error analysis involves the need for an innerproduct in the linear space M_{nn} of complex $n \times n$ matrices. Hence we define $(A, B)_E := \sum_{\ell, m=1}^n A_{\ell m} \bar{B}_{\ell m}$, $A, B \in M_{nn}$; this product induces the many times used Euclidean (Frobenius) norm $\|A\|_E := (A, A)_E^{1/2}$. In the product space $M_{nn} \times C^n$ we consider the innerproduct

$$((A, x), (B, y))_\alpha := (A, B)_E + \alpha(x, y)_2, (A, x), (B, y) \in M_{nn} \times C^n,$$

where $\alpha > 0$.

THEOREM 4.2. Let be given $A \in M_{nn}$, $x \in C^n \setminus \{0\}$, $\lambda \in C^n$ and $\alpha > 0$. The mappings $f, P : M_{nn} \times C^n \rightarrow C^n$ are defined by

$$\left\{ \begin{array}{l} f(H, h) := (A+H)(x+h) - \lambda(x+h) \\ P(H, h) := (A-\lambda I)h + Hx. \end{array} \right. \quad (4.6)$$

$$(H, h) \in M_{nn} \times C^n, \quad (4.7)$$

With respect to the innerproducts $(,)_\alpha$ in $M_{nn} \times C^n$ and the usual Euclidean innerproduct in C^n

$$PP^* = \|x\|^2 I + \frac{1}{\alpha}(A-\lambda I)(A-\lambda I)^* . \quad (4.8)$$

Df satisfies a Lipschitz condition with Lipschitz constant $L = \frac{2}{\alpha}$. If

- (i) $\|(PP^*)^{-1}\|_2^{\frac{1}{2}} = \mu_n^{-1} \leq \mu^{-1}$, μ_n being the smallest singular value of P ;
- (ii) $(r^*(PP^*)^{-1}r)^{\frac{1}{2}} = \tilde{\gamma} \leq \gamma$, where $r := f(0,0) = Ax - \lambda x$;
- (iii) $\kappa := \gamma \mu^{-1} L < \frac{1}{2}$,

then the distance \hat{d} of $(0,0)$ to $\hat{S} := \{(H, h) | f(H, h) = 0\}$ satisfies the inequalities

$$\rho_1 := \rho_3 \left(1 + \frac{L^2 \rho_3^2}{(\mu - L\rho_3)^2}\right)^{-\frac{1}{2}} < \hat{d} < \rho$$

where

$$\rho = \mu(1 - \sqrt{1 - 2\kappa})/L, \quad \rho_3 = \mu(\sqrt{1 + 2L\tilde{\gamma}\mu^{-1}} - 1)/L.$$

PROOF. With respect to the innerproducts $(,)_\alpha$ in $M_{nn} \times C^n$ and $(,)_2$ in C^n

we derive for the adjoint $P^*: C^n \rightarrow M_{nn} \times C^n$ of $P = Df(0,0)$:

$$\begin{aligned} ((H,h), P^*k)_\alpha &= (P(H,h), k)_2 = ((A-\lambda I)h + Hx, k)_2 \\ &= (h, (A-\lambda I)^*k)_2 + (H, kx^*)_E = ((H,h), (kx^*, \frac{1}{\alpha}(A-\lambda I)^*k))_\alpha. \end{aligned}$$

So for each $k \in C^n$ we have

$$P^*k = (kx^*, \frac{1}{\alpha}(A-\lambda I)^*k). \quad (4.9)$$

With (4.7) we find for $PP^*: C^n \rightarrow C^n$:

$$PP^*k = (\frac{1}{\alpha}(A-\lambda I)(A-\lambda I)^* + \|x\|_2^2 I)k.$$

The Hermitian matrix PP^* is positive definite, thus P is surjective. Now, we compute a Lipschitz constant for Df . Evidently, $Df(E,e)(H,h) = (A+E-\lambda I)h + H(x+e)$. Thus,

$$\begin{aligned} \|Df(E_1, e_1) - Df(E_2, e_2)(H,h)\|_2^2 &= \|(E_1-E_2)h + H(e_1-e_2)\|_2^2 \\ &\leq 2\|H\|_2^2 \|e_1-e_2\|_2^2 + 2\|E_1-E_2\|_2^2 \|h\|_2^2 \\ &\leq \frac{2}{\alpha}(\|E_1-E_2\|_2^2 + \alpha\|e_1-e_2\|_2^2)(\|H\|_2^2 + \alpha\|h\|_2^2) \\ &\leq \frac{2}{\alpha} \|(E_1-E_2, e_1-e_2)\|_\alpha^2 \cdot \|(H,h)\|_\alpha^2. \end{aligned}$$

Consequently,

$$\|Df(E_1, e_1) - Df(E_2, e_2)\| \leq \sqrt{\frac{2}{\alpha}} \|(E_1-E_2, e_1-e_2)\|_\alpha.$$

So for the Lipschitz constant L of derivative Df we have found $L \leq \sqrt{\frac{2}{\alpha}}$. With this bound for L we are able to apply theorem 2.1; that gives the upper bound ρ and the lower bound ρ_1 for the necessary perturbation (H,h) such that $f(H,h) = 0$. \square

This application of the minimization process in the normal space of a level surface of f at $(0,0)$ can be applied for the construction of vector $x+h$ nearby to x and a matrix $A+H$, nearby to A , with a prescribed eigen-

value. A small weight α in the innerproduct $(,)_\alpha$ effects in the minimization result the term $\|H\|_E^2$ in $\|(H,h)\|_\alpha^2$ to be small relatively $\|h\|_2^2$.

In our experiments the modified method of Newton, in formula

$$(H_k, h_k) := (H_{k-1}, h_{k-1}) - P^*(PP^*)^{-1} f(H_{k-1}, h_{k-1}), \quad k \in \mathbb{N},$$

with starting value $(H_0, h_0) = (0, 0) \in M_{nn} \times C^n$, has been implemented in the following way.

Rowwise the matrix H_k is set in a linear n^2 array. So the $n \times (n^2+n)$ matrix P equals the Kroneckerproduct $I \otimes x^T$, bordered by $A-\lambda I$. Then the matrix representation of the adjoint P^* of P ($(,)_\alpha$) equals, as follows from (4.9),

$$\begin{pmatrix} I \otimes \bar{x} \\ \dots \\ \alpha^{-1} \overline{A-\lambda I}^T \end{pmatrix}$$

The Newton correction $P^*(PP^*)^{-1} f(H_{k-1}, h_{k-1})$ has been obtained from the QR-decomposition of the appropriately adapted P^* :

$$QR = \begin{pmatrix} I \otimes \bar{x} \\ \dots \\ \alpha^{-1/2} \overline{A-\lambda I}^T \end{pmatrix}$$

Then

$$PP^* = \bar{R}^T \bar{Q}^T QR.$$

Consequently,

$$(H_k, h_k) = (H_{k-1}, h_{k-1}) - \bar{Q}\bar{R}^{-T} f(H_{k-1}, h_{k-1}), \quad k \in \mathbb{N}.$$

So, after the initial QR-decomposition, in each step one has to solve a lower triangular linear system with n unknowns.

As a test we used this algorithm on the Frank matrix F of order 11, (λ, x) being, as in the preceding example, a single-precision approximation of the eigenpair corresponding with the smallest eigenvalue of F . Single

precision words on ICL 1902 are of 14 hexadecimals and double-precision words (long reals) are of 28 hexadecimals, these correspond to 16.8 and 33.7 decimal digits respectively. In this example we have chosen $\alpha = 10^{-12}$. The smallest eigenvalue is, as mentioned above, ill-conditioned, so λ is fairly inaccurate, but $f(H_0, h_0) = x = Ax - \lambda x$ was of the order of the noise level of the computer, i.e. 2^{-56} .

In table 2 and 3 we give an informative selection of the numerical results.

Table 2	
$\ f(H_3, h_3)\ $	$= 5.099 * 10^{-33}$
$\ H_3\ _E$	$= 1.423 * 10^{-12}$
$\ h_3\ _2$	$= 3.562 * 10^{-6}$
$\ (H_3, h_3)\ _{10^{-12}}$	$= 3.835 * 10^{-12}$

Table 3	
(i, j)	$(H_3)_{ij}$
(2, 11)	$7.151 * 10^{-13}$
(2, 10)	$-6.904 * 10^{-13}$
(1, 11)	$-6.522 * 10^{-13}$
(1, 10)	$6.296 * 10^{-13}$
(2, 9)	$3.209 * 10^{-13}$
(1, 9)	$-2.926 * 10^{-13}$
(2, 8)	$-0.953 * 10^{-13}$
(1, 8)	$0.869 * 10^{-13}$
(3, 11)	$-0.672 * 10^{-13}$
(3, 10)	$0.649 * 10^{-13}$
(8, 1)	$2.241 * 10^{-27}$!

Extreme elements of H_3

The smallness of this feasible perturbation H_3 bases on the ill-condition of the smallest eigenvalue. As the individual condition number 10^6 effects that a perturbation of order 10^{-12} results in an error of order 10^{-6} [19], here we see that a forced error of order 10^{-6} can be explained by a perturbation of order 10^{-12} .

The described method can be generalized by construct a matrix $A+H$, nearby to a given matrix A , such that the spectrum of $A+H$ differs in a prescribed way from that of A . In a next memorandum, we report about that problem.

5. REFERENCES

1. d'Allessandro, P.: Structural Properties of Bilinear Discrete-time systems, *Richerche di Automatica*, 3, 1972.
2. Anselone, P.M. and L.B. Rall: The Solution of Characteristic Value Problems by Newton's Method, *Num. Math.*, 11, 1968, pp. 38-45.
3. Antosiewicz, H.A.: Newton's Method and Boundary Value Problems, *J. Comp. Syst. Sc.*, 2, 1968, pp. 177-202.
4. Aubin, J.P.: *Applied Functional Analysis*, Wiley, New York, xv + 423 pp., 1979.
5. Bellman, R.E. and R.E. Kalaba: *Quasilinearization and Nonlinear Boundary Value Problems*, Elsevier, New York, 1965.
6. Bruni, C., et al.: Bilinear Systems: An Appealing Class of "Nearly Linear" Systems in Theory and Applications, *IEEE, AC*, 19, 1974, pp. 334-348.
7. Coddington, A. and N. Levinson: *Theory of Ordinary Differential Equations*, McGraw Hill, New York, 1955.
8. Gragg, W.B. and R.A. Tapia: Optimal Error Bounds for the Newton Kantorovich Theorem, *SIAM J. Num. An.*, 11, 1974, pp. 10-13.
9. Kantorovich, L.: *Functional Analysis and Applied Mathematics*, *Uspehi Mat. Nauk* 3, 1948, pp. 89-185; transl. by C. Benster *Nat. Bur. Standards Report 1509*, Washington D.C., 1952.
10. Kato, T.: *Perturbation Theory for Linear Operators*, Springer, Berlin, 1966.
11. Mohler, R.R.: *Bilinear Control Processes with Applications to Engineering, Ecology and Medicine*, Acad. Press, New York, 1973.

12. Mohler, R.R. and A. Ruberti: Theory and Applications of Variable Structure Systems, Acad. Press, New York, 1972, xii + 232 pp.
13. Ortega, J.M.: The Newton-Kantorovich Theorem, Amer. Math. Monthly, 75, 1968, pp. 658-660.
14. Paardekooper, M.H.C.: A Newton-Kantorovich Method to Analyse Errors in Least Squares Problems, Proceedings ESEM 1980, Athens.
15. Paardekooper, M.H.C.: The Distribution of Errors in Linear Models with Time Varying Parameters and Errors in the Observables; Reeks "Ter Discussie", 80.11, Tilburg University.
16. Symm, H.J. and J.H. Wilkinson: Realistic Error Bounds for a Simple Eigenvalue and its Associated Eigenvector, Num. Math., 35, 1980, pp. 113-126.
17. Tapia, R.A.: The Differentiation and Integration of Nonlinear Operators, in Nonlinear Functional Analysis and Applications, L.B. Rall, e d, Academic Press, New York, 1971, pp. 45-103.
18. Taylor, A.E. and D.C. Lay: Introduction to Functional Analysis, 2nd Edition, Wiley, New York, xii + 467 pp., 1980.
19. Wilkinson, J.H.: The Algebraic Eigenvalue Problem, Oxford Un. Press, London, 1965, xviii + 662 pp.
20. Wouk, A.: A Course of Applied Functional Analysis, Wiley, New York, xvii + 433 pp., 1979.
21. Granger, C.W.J. and A. Andersen: On Introduction to Bilinear Time Series Models, Van den Hoek & Ruprecht, 1978, 91 pp.

Bibliotheek K. U. Brabant



17 000 01059860 6