MATHEMATICS OF COMPUTATION
VOLUME 38, NUMBER 157
JANUARY 1982

Numerical Stability of the Halley-Iteration for the Solution of a System of Nonlinear Equations

By Annie A. M. Cuyt*

Abstract. Let $F\colon \mathbf{R}^q \to \mathbf{R}^q$ and x^* a simple root in $\mathbf{R}^q$ of the system of nonlinear equations $F(x) = 0$.

Abstract Padé approximants (APA) and abstract Rational approximants (ARA) for the operator F have been introduced in [2] and [3]. The adjective "abstract" refers to the use of abstract polynomials [5] for the construction of the rational operators.

The APA and ARA have been used for the solution of a system of nonlinear equations in [4]. Of particular interest was the following third order iterative procedure:

$$x_{i+1} = x_i + \frac{a_i^2}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2},$$

with F_i' the 1st Fréchet-derivative of F in x_i, $a_i = -F_i'^{-1}F_i$ the Newton-correction where $F_i = F(x_i)$, F_i'' the 2nd Fréchet-derivative of F in x_i where $F_i''a_i^2$ is the bilinear operator F_i'' evaluated in (a_i, a_i), and componentwise multiplication and division in $\mathbf{R}^q$. For $q = 1$ this technique is known as the Halley-iteration [6, p. 91]. In this paper the numerical stability [7] of the Halley-iteration for the case $q > 1$ is investigated and illustrated by a numerical example.

1. Numerical Stability of Iterations. We consider the numerical solution of the equation

$$F(x) = 0 \tag{1}$$

with $F\colon \mathbf{R}^q \to \mathbf{R}^q\colon x \to F(x)$, abstract analytic in 0 [5]. Assume that (1) has a simple root x^*.

We briefly repeat the definition of condition-number given by Woźniakowski [7]. The condition-number should measure the sensitivity of the solution (output) with respect to changes in the data (input). We assume that F depends parametrically on a vector $d \in \mathbf{R}^p$, called data vector

$$F(x) = F(x; d),$$

and instead of the exact value $F(x; d)$ we only have the computed value $\mathrm{fl}(F(x; d))$ in t digit floating-point binary arithmetic. At best we can expect that $\mathrm{fl}(F(x; d))$ is the exact value of a slightly perturbed operator at slightly perturbed data

$$\mathrm{fl}(F(x; d)) = (I + \Delta F)F(x + \Delta x; d + \Delta d), \tag{2}$$

where I is the $q \times q$ unit-matrix and

$$\|\Delta x\| \leqslant C_1\rho\|x\|, \qquad \|\Delta d\| \leqslant C_2\rho\|d\|,$$
$$\|\Delta F\| \leqslant C_3\rho \quad (\Delta F \text{ a } q \times q \text{ matrix}),$$

Received May 27, 1981.

1980 *Mathematics Subject Classification.* Primary 65G05.

* Aspirant of the Belgian National Fund for Scientific Research (NFWO).

0025-5718/82/0000-0478/$03.25

for constants C_1, C_2, C_3 (only depending on the dimensions of the problem) and with $\rho = 2^{-t}$ the relative computer precision [8]. By introducing the Landau-symbol O, we could also write

$$\Delta x = O(\rho), \quad \Delta d = O(\rho), \quad \Delta F = O(\rho),$$

where the constants in the Landau-notation depend on x, d and the dimensions. We will always, for a given F, define the data vector so that (2) holds and so that the condition number (see Definition 1.1) is minimized. Let $\mathrm{fl}(d)$ denote the t digit binary representation of the vector d in floating-point arithmetic

$$\|\mathrm{fl}(d) - d\| \leqslant C\rho\|d\|, \quad \text{i.e. } \mathrm{fl}(d) - d = O(\rho).$$

Since d is represented by $\mathrm{fl}(d)$, we solve in fact $F(x; \mathrm{fl}(d)) = 0$ instead of $F(x) = 0$, independent of the method used to solve (1). Let F'_x and F'_d denote the partial Fréchet-derivatives of F, respectively with respect to x and d.

Now $F(x; \mathrm{fl}(d)) = 0$ has a root $\widetilde{x^*}$ in the neighborhood of x^* and $\widetilde{x^*} - x^* = O(\rho)$ if t is sufficiently large; thus,

$$\begin{aligned}\widetilde{x^*} - x^* &= -F'_x(x^*; d)^{-1}F'_d(x^*; d)(\mathrm{fl}(d) - d)\\ &\qquad + \text{higher order terms in } \widetilde{x^*} - x^* \text{ and } \mathrm{fl}(d) - d\\ &= -F'_x(x^*; d)^{-1}F'_d(x^*; d)(\mathrm{fl}(d) - d) + O(\rho^2),\end{aligned}$$

where the constant in the Landau-notation depends on x^*, d and F.

For $x^* \neq 0$: $\|\widetilde{x^*} - x^*\|/\|x^*\| \leqslant \|F'_x(x^*; d)^{-1}F'_d(x^*; d)\|C\rho\|d\|/\|x^*\| + O(\rho^2)$.

Definition 1.1. $\mathrm{Cond}(F; d) = \|F'_x(x^*; d)^{-1}F'_d(x^*; d)\| \cdot \|d\|/\|x^*\|$ is called the *condition number of F* with respect to the data vector d.

A problem is ill-conditioned if $\mathrm{cond}(F; d) \gg 1$.

Let us now suppose that $F(x; d) = 0$ is solved by an iterative procedure $\Phi(x_i, F)$, where Φ can use several $F_i^{(j)}$, the jth Fréchet-derivative of F at x_i (if $j = 1$ or 2, a single or double prime is used instead of the superscript j). If $\{x_i\}$ is the sequence of successive approximations of x^*, we can at best expect x_i to be the representation of a computed value for $\widetilde{x^*}$,

$$\|x_i - \widetilde{x^*}\| \leqslant K\rho\|\widetilde{x^*}\|.$$

So

$$\begin{aligned}\|x_i - x^*\| &\leqslant \|x_i - \widetilde{x^*}\| + \|\widetilde{x^*} - x^*\| \leqslant K\rho\|\widetilde{x^*}\| + C\rho\,\mathrm{cond}(F; d)\cdot\|x^*\| + O(\rho^2)\\ &\leqslant K\rho(\|\widetilde{x^*} - x^*\| + \|x^*\|) + C\rho\,\mathrm{cond}(F; d)\cdot\|x^*\| + O(\rho^2)\\ &\leqslant [K\rho + C\rho\,\mathrm{cond}(F; d)]\cdot\|x^*\| + O(\rho^2).\end{aligned}$$

Definition 1.2. An iteration Φ is called *numerically stable* if

$$\lim_{i\to\infty}\|x_i - x^*\| \leqslant \rho\cdot\|x^*\|\cdot(C\,\mathrm{cond}(F; d) + K) + O(\rho^2),$$

where the constants C and K depend on x^*, d and F.

In practice we often want to find an approximation x_i such that $\|x_i - x^*\| \leqslant \varepsilon\cdot\|x^*\|$. This is possible if the problem is sufficiently well-conditioned, i.e., $\rho\,\mathrm{cond}(F; d) = O(\varepsilon)$. In floating-point arithmetic we have

$$x_{i+1} = \Phi(x_i, F) + \xi_i, \quad \text{where } \xi_i = \mathrm{fl}(\Phi(x_i, F)) - \Phi(x_i, F).$$

THEOREM 1.1. *A convergent iterative procedure* $\Phi(x_i, F)$, *i.e.*

$$\lim_{i\to\infty} \|\Phi(x_i, F) - x^*\| = 0,$$

is numerically stable if $\lim_{i\to\infty}\|\xi_i\| \leqslant \rho\|x^*\|(C \operatorname{cond}(F; d) + K) + O(\rho^2)$.

Proof. We simply verify the definition.

$$\lim_{i\to\infty} \|x_i - x^*\| \leqslant \lim_{i\to\infty} \left[\|\Phi(x_{i-1}, F) - x^*\| + \|\xi_{i-1}\| \right]$$
$$= \lim_{i\to\infty} \|\xi_{i-1}\| \leqslant \rho\|x^*\|(C \operatorname{cond}(F; d) + K) + O(\rho^2).$$

2. Abstract Padé Approximants (APA) and Abstract Rational Approximants (ARA) for the Solution of a System of Nonlinear Equations. Let x_i be the ith approximant of the root x^* in the iterative process, $y_i = F(x_i)$ and the Newton-correction $a_i = -F_i'^{-1}F_i$. Using the Inversion Theorem [1, p. 381] we can see that

$$x^* = x_i + a_i - \tfrac{1}{2}F_i'^{-1}F_i''a_i^2 + O(a_i^3), \tag{3}$$

where $F_i''a_i^2$ is the bilinear operator F_i'' evaluated on (a_i, a_i). The Newton-iteration results from approximating the series in (3) by its first two terms, i.e., the (1, 0)-APA [2].

In [7] Woźniakowski proves numerical stability of the Newton-iteration under a natural assumption on the computed evaluation of F.

THEOREM 2.1. *If*

(a) $\mathrm{fl}(F(x_i; d)) = (I + \Delta F_i)F(x_i + \Delta x_i; d + \Delta d_i) = F(x_i) + \delta F_i$, *with*

$$\delta F_i = \Delta F_i F(x_i) + F_x'(x_i)\Delta x_i + F_d'(x_i)\Delta d_i + O(\rho^2),$$

(b) $\mathrm{fl}(F'(x_i; d)) = F'(x_i) + \delta F_i'$, *with* $\delta F_i' = O(\rho)$,

(c) *the computed correction* $\mathrm{fl}(a_i)$ *is the exact solution of a perturbed linear system*

$$(F'(x_i) + \delta F_i' + E_i)\mathrm{fl}(a_i) = -F(x_i) - \delta F_i \quad \textit{with } E_i = O(\rho),$$

then the Newton-iteration is numerically stable.

Proof. In [7].

Another way to approximate x^* is to use the (1, 1)-ARA [2] for the power series (3), i.e.

$$x_{i+1} = x_i + \frac{a_i^2}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2}, \tag{4}$$

where multiplication and division of the vectors in $\mathbf{R}^q$ in the numerator and denominator of (4) are componentwise. For $q = 1$ the iteration (4) is the well-known Halley-iteration. We will also use the name Halley-iteration for the case $q \geqslant 1$. We will now prove numerical stability of this iteration under assumptions similar to the assumptions for the Newton-iteration. We will also assume that the divisions in (4) are such that

$$\left(\frac{1}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2}\right)^j O(\|a_i\|^{j-k}\rho^{k+l}) = O(\rho^l). \tag{5}$$

Condition (5) takes care of the fact that the denominator of the correction-term in (4) does not become too small in comparison with $O(\|a_i\|^{j-k}\rho^k)$.

The assumption of (5) is a natural generalization of the following relations:

$$\text{for } q = 1,\ \lim_{i\to\infty} \frac{a_i}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2} = 1,$$

(5a)
$$\text{and so } \exists L \in \mathbf{N} \ni \forall i \geqslant L: \left| \frac{a_i}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2} \right| \leqslant 1 + D$$

(case $j = 1, k = 0, l = 0$) with $D \in \mathbf{R}_0^+$,

in a convergent process (4): $\lim_{i\to\infty} \|x^* - x_i\| = 0$, and thus

$$\lim_{i\to\infty} a_i = 0, \quad \text{i.e. } \exists M \in \mathbf{N} \ni \forall i \geqslant M: a_i = O(\rho),$$

and so $\forall i \geqslant M$: $a_i^2 = O(\|a_i\|\rho)$; also

$$\lim_{i\to\infty} \frac{a_i^2}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2} = 0, \text{ i.e.}$$

(5b)
$$\exists N \in \mathbf{N} \ni \forall i \geqslant N: \frac{a_i^2}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2} = O(\rho),$$

$$\text{and so } \forall i \geqslant \max(N, M): \frac{a_i^2}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2} = \frac{1}{a_i + \frac{1}{2}F_i'^{-1}F_i''a_i^2} O(\|a_i\|\rho) = O(\rho)$$

(case $j = 1, k = 0, l = 1$).

THEOREM 2.2. *If*

(a) $\mathrm{fl}(F(x_i; d)) = (I + \Delta F_i)F(x_i + \Delta x_i; d + \Delta d_i) = F(x_i) + \delta F_i$ *with*

$$\delta F_i = \Delta F_i F(x_i) + F_x'(x_i)\Delta x_i + F_d'(x_i)\Delta d_i + O(\rho^2),$$

(b) $\mathrm{fl}(F'(x_i; d)) = F'(x_i) + \delta F_i'$ *with* $\delta F_i' = O(\rho)$,

(c) $\mathrm{fl}(F''(x_i; d)) = F''(x_i) + \delta F_i''$ *with* $\delta F_i'' = O(\rho)$,

(d) *the computed correction* $\mathrm{fl}(a_i)$ *is the exact solution of a perturbed linear system*

$$(F'(x_i) + \delta F_i' + E_{i,1})\mathrm{fl}(a_i) = -F(x_i) - \delta F_i \quad \text{with } E_{i,1} = O(\rho),$$

(e) *analogously*,

$$(F'(x_i) + \delta F_i' + E_{i,2})\mathrm{fl}(b_i) = (F''(x_i) + \delta F_i'')\mathrm{fl}(a_i)^2$$

$$\text{with } E_{i,2} = O(\rho) \text{ and } b_i = F_i'^{-1}F_i''a_i^2,$$

and (5) *holds, then the iteration* (4) *is numerically stable.*

Proof. Let $F'(x_i) + \delta F_i' + E_{i,1} = F'(x_i)(I + H_{i,1})$, where

$$H_{i,1} = F'(x_i)^{-1}\{\delta F_i' + E_{i,1}\} = O(\rho)$$

because of (b) and (d). So for small ρ,

$$(I + H_{i,1})^{-1} = I - H_{i,1} + O(\rho^2).$$

Thus

$$\text{fl}(a_i) = (I - H_{i,1})F_i'^{-1}(-F_i - \delta F_i). \tag{6}$$

Analogously

$$\text{fl}(b_i) = (I - H_{i,2})F_i'^{-1}(F_i'' + \delta F_i'')\text{fl}(a_i)^2 \quad \text{with } H_{i,2} = O(\rho).$$

Now

$$\begin{aligned}(F_i'' + \delta F_i'')\text{fl}(a_i)^2 &= (F_i'' + \delta F_i'')\big[(I - H_{i,1})F_i'^{-1}(-F_i - \delta F_i)\big]^2\\ &= (F_i'' + \delta F_i'')a_i^2 + 2(F_i'' + \delta F_i'')\big(F_i'^{-1}F_i, F_i'^{-1}\delta F_i - H_{i,1}F_i'^{-1}F_i\big) + O(\rho^2)\\ &= (F_i'' + \delta F_i'')a_i^2 - 2F_i''\big(a_i, F_i'^{-1}\delta F_i - H_{i,1}F_i'^{-1}F_i\big) + O(\rho^2).\end{aligned}$$

Thus

$$\begin{aligned}\text{fl}(b_i) = {}& F_i'^{-1}(F_i'' + \delta F_i'')a_i^2 - 2F_i'^{-1}F_i''\big(a_i, F_i'^{-1}\delta F_i - H_{i,1}F_i'^{-1}F_i\big)\\ &- H_{i,2}F_i'^{-1}F_i''a_i^2 + O(\rho^2).\end{aligned}$$

A computed approximation x_{i+1} satisfies

$$x_{i+1} = (I + \delta I_{i,1})\left[x_i + (I + \delta I_{i,2})\frac{\text{fl}(a_i)^2}{\text{fl}(a_i) + \frac{1}{2}\text{fl}(b_i)}\right],$$

where $\delta I_{i,1}$ and $\delta I_{i,2}$ are diagonal matrices and $\delta I_{i,1} = O(\rho)$ and $\delta I_{i,2} = O(\rho)$. So

$$x_{i+1} = (I + \delta I_{i,1})\left[x_i + (I + \delta I_{i,2})\frac{a_i^2 - 2a_i \cdot \big(F_i'^{-1}\delta F_i + H_{i,1}a_i\big) + O(\rho^2)}{a_i + \frac{1}{2}b_i - \delta a_i + O(\rho^2)}\right],$$

where

$$\begin{aligned}\delta a_i = {}& F_i'^{-1}\delta F_i + H_{i,1}a_i - \tfrac{1}{2}F_i'^{-1}\delta F_i''a_i^2\\ &+ \tfrac{1}{2}H_{i,2}F_i'^{-1}F_i''a_i^2 + F_i'^{-1}F_i''\big(a_i, F_i'^{-1}\delta F_i - H_{i,1}F_i'^{-1}F_i\big).\end{aligned}$$

Using (6), we find

$$\text{fl}(a_i) - a_i + H_{i,1}a_i - H_{i,1}F_i'^{-1}\delta F_i = -F_i'^{-1}\delta F_i,$$

and thus, for positive constants D_1 and D_2,

$$\|F_i'^{-1}\delta F_i\| \leqslant D_2\rho\|a_i\| \quad \text{since } \|\text{fl}(a_i) - a_i\| \leqslant D_1\rho\|a_i\|$$

and

$$\|F_i'^{-1}\| \cdot \|F_i\| \leqslant \|F_i'^{-1}\| \cdot \|F_i'\| \cdot \|a_i\|.$$

Thus

$$x_{i+1} = (I + \delta I_{i,1})\left[x_i + \frac{a_i^2 - 2a_i\big(F_i'^{-1}\delta F_i + H_{i,1}a_i\big) + \delta I_{i,2}a_i^2 + O\big(\rho^2\|a_i\|^2\big)}{a_i + \frac{1}{2}b_i - \delta a_i + O(\rho^2)}\right],$$

where $\delta I_{i,2}a_i^2$ is the linear operator $\delta I_{i,2}$ evaluated in a_i^2 (componentwise square of the vector a_i). So

$$x_{i+1} = (I + \delta I_{i,1})\left[x_i + \frac{a_i^2 - 2a_i\big(F_i'^{-1}\delta F_i + H_{i,1}a_i\big) + \delta I_{i,2}a_i^2 + O\big(\rho^2\|a_i\|^2\big)}{a_i + \frac{1}{2}b_i} \cdot c_i\right],$$

with

$$c_i = 1 + \frac{1}{a_i + \frac{1}{2}b_i}(\delta a_i + O(\rho^2)) + \left(\frac{1}{a_i + \frac{1}{2}b_i}\right)^2 O\left(\|a_i\|^{2-k}\rho^{k+2}, k = 0, 1, 2\right)$$

since $\delta a_i = O(\rho\|a_i\|)$; in c_i we have used the notation 1 for the unit vector $(1, \ldots, 1)$.

Using (5), we conclude

$$\left(\frac{1}{a_i + \frac{1}{2}b_i}\right)^2 O\left(\|a_i\|^{2-k}\rho^{k+2}, k = 0, 1, 2\right) = O(\rho^2).$$

For $\xi_i = x_{i+1} - \Phi(x_i, F)$, we have

$$\begin{aligned}\xi_i = \delta I_{i,1}x_i &+ \frac{a_i^2}{a_i + \frac{1}{2}b_i}(c_i - 1)\\ &+ \frac{-2a_i\left(F_i'^{-1}\delta F_i + H_{i,1}a_i\right) + \delta I_{i,2}a_i^2 + O\left(\rho^2\|a_i\|^2\right)}{a_i + \frac{1}{2}b_i}\cdot c_i\\ &+ \delta I_{i,1}\frac{a_i^2}{a_i + \frac{1}{2}b_i}\cdot c_i + O(\rho^2).\end{aligned}$$

So

$$\begin{aligned}\xi_i = \delta I_{i,1}x_i &+ \left(\frac{1}{a_i + \frac{1}{2}b_i}\right)^2 O\left(\rho\|a_i\|^3, \rho^2\|a_i\|^2\right) + \frac{1}{a_i + \frac{1}{2}b_i}O\left(\rho^2\|a_i\|^2\right)\\ &+ \frac{1}{a_i + \frac{1}{2}b_i}\left(-2a_iF_i'^{-1}\delta F_i + O\left(\rho\|a_i\|^2, \rho^2\|a_i\|^2\right)\right)\cdot(1 + O(\rho))\\ &+ O(\rho^2).\end{aligned}$$

Thus

$$\|\xi_i\| \leqslant k_1\rho\|x_i\| + k_2\rho\|a_i\| + \left\|\frac{-2a_i}{a_i + \frac{1}{2}b_i}F_i'^{-1}\delta F_i\right\| + O(\rho^2),$$

and since

$$\begin{aligned}\frac{-2a_i}{a_i + \frac{1}{2}b_i}F_i'^{-1}\delta F_i &= \frac{-2a_i}{a_i + \frac{1}{2}b_i}F_i'^{-1}\left(\Delta F_iF(x_i) + F_i'\Delta x_i + F_d'\Delta d_i + O(\rho^2)\right)\\ &= \frac{1}{a_i + \frac{1}{2}b_i}O(\rho\|a_i\|)F(x_i) - \frac{2a_i}{a_i + \frac{1}{2}b_i}\Delta x_i\\ &\quad - \frac{2a_i}{a_i + \frac{1}{2}b_i}F_i'^{-1}F_d'\Delta d_i + \frac{1}{a_i + \frac{1}{2}b_i}O\left(\rho^2\|a_i\|\right),\end{aligned}$$

we find that

$$\lim_{i\to\infty}\|\xi_i\| \leqslant \rho\|x^*\|(K + C\,\mathrm{cond}(F; d)) + O(\rho^2)$$

for $\lim_{i\to\infty} a_i = 0 = \lim_{i\to\infty} F(x_i)$ in a convergent process and $a_i\Delta x_i = O(\rho\|a_i\|)$ and $a_iF_i'^{-1}F_d'\Delta d_i = O(\rho\|a_i\|)$.

3. Numerical Example. Consider the following operator:

$$F\colon \mathbf{R}^2 \to \mathbf{R}^2\colon (x, y) \to \begin{pmatrix} e^{-x+y} - d_1 \\ e^{-x-y} - d_2 \end{pmatrix} \quad \text{with } d_1 > 0 \text{ and } d_2 > 0.$$

The operator F has a simple root $x^* = (-\frac{1}{2}\ln(d_1 d_2), \frac{1}{2}\ln(d_1/d_2))$. Clearly

$$d = (d_1, d_2)$$

is the data vector. Now

$$\mathrm{fl}(F(x, y; d)) = \begin{bmatrix} \left[(1+\varepsilon_1)e^{(-x-\Delta' x+y+\Delta' y)(1+\theta_1)} - (d_1 + \Delta_1' d)\right](1+\kappa_1) \\ \left[(1+\varepsilon_2)e^{(-x-\Delta' x-y-\Delta' y)(1+\theta_2)} - (d_2 + \Delta_2' d)\right](1+\kappa_2) \end{bmatrix},$$

where $\mathrm{fl}(x) = x + \Delta' x$, $\mathrm{fl}(y) = y + \Delta' y$, $\mathrm{fl}(d_1) = d_1 + \Delta_1' d$, $\mathrm{fl}(d_2) = d_2 + \Delta_2' d$, θ_1 is caused by $-\mathrm{fl}(x) + \mathrm{fl}(y)$, θ_2 is caused by $-\mathrm{fl}(x) - \mathrm{fl}(y)$, ε_i are caused by the exponential evaluations ($i = 1, 2$), κ_i are caused by the subtraction of $\mathrm{fl}(d_i)$ ($i = 1, 2$).

One can rewrite $\mathrm{fl}(F(x, y; d)) = (I + \Delta F)F(x + \Delta x, y + \Delta y; d + \Delta d)$ with

$$\Delta x = x\theta_1 + \Delta' x(1 + \theta_1), \quad \Delta y = y\theta_1 + \Delta' y(1 + \theta_1), \quad \Delta d = (\Delta_1 d, \Delta_2 d),$$

$$\Delta_1 d = \frac{\Delta_1' d - \varepsilon_1 d_1}{1 + \varepsilon_1},$$

$$\Delta_2 d = \frac{\Delta_2' d - \varepsilon_2 d_2}{1 + \varepsilon_2} + \frac{d_2 + \Delta_2' d}{1 + \varepsilon_2}\left(e^{(x+\Delta' x+y+\Delta' y)(\theta_2 - \theta_1)} - 1\right),$$

$$\Delta F = \begin{pmatrix} (1+\varepsilon_1)(1+\kappa_1) - 1 & 0 \\ 0 & (1+\varepsilon_2)(1+\kappa_2)e^{(x+\Delta' x+y+\Delta' y)(\theta_1 - \theta_2)} - 1 \end{pmatrix}.$$

The inverse of the Jacobian matrix in the root x^* is

$$\frac{1}{2(d_1 \cdot d_2)}\begin{pmatrix} -d_2 & -d_1 \\ d_2 & -d_1 \end{pmatrix} \quad \text{and} \quad F_d' = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The condition number of F with respect to the data vector d is

$$\|F_x'(x^*; d)^{-1}\| \cdot \frac{\|(d_1, d_2)\|}{\|x^*\|}.$$

Using the Schur-norm $\|A\| = \sqrt{\sum_{i,j} a_{ij}^2}$ of a matrix $A = (a_{ij})$ and the l_2-norm $\|a\| = \sqrt{\sum_i a_i^2}$ of a vector $a = (a_i)$, the condition number is

$$\frac{d_1^2 + d_2^2}{\sqrt{2}\, d_1 \cdot d_2 \cdot \|x^*\|}.$$

Putting $d_1 = d = d_2$, the root $x^* = (-\ln d, 0)$ and the condition number is $\sqrt{2}/|\ln d|$. The problem is extremely well-conditioned if $\mathrm{cond}(F; d) \leqslant 1$, i.e.,

$$d \in \left]-\infty, e^{-\sqrt{2}}\right] \cup \left[e^{\sqrt{2}}, +\infty\right[.$$

The problem is very ill-conditioned if $d = e^{\varepsilon}$ with ε very small. We will now check some of the conditions of Theorem 2.2. We already know $\mathrm{fl}(F(x, y; d)) = (I + \Delta F)F(x + \Delta x, y + \Delta y; d + \Delta d)$.

Now

$$\mathrm{fl}(F'(x, y; d)) = \mathrm{fl}\begin{pmatrix} -e^{-x+y} & e^{-x+y} \\ -e^{-x-y} & -e^{-x-y} \end{pmatrix},$$

where

$$\begin{aligned} \mathrm{fl}(e^{-x+y}) &= (1+\varepsilon_1)e^{(-x-\Delta'x+y+\Delta'y)(1+\theta_1)} = (1+\varepsilon_1)e^{-x+y}e^{-\Delta x+\Delta y} \\ &= e^{-x+y}\left[1+\varepsilon_1+(1+\varepsilon_1)(e^{-\Delta x+\Delta y}-1)\right], \\ \mathrm{fl}(e^{-x-y}) &= (1+\varepsilon_2)e^{(-x-\Delta'x-y-\Delta'y)(1+\theta_2)} \\ &= (1+\varepsilon_2)e^{-x-y}e^{-\Delta x-\Delta y}e^{(x+\Delta'x+y+\Delta'y)(\theta_1-\theta_2)} \\ &= e^{-x-y}\left[1+\varepsilon_2+(1+\varepsilon_2)(e^{-\Delta x-\Delta y}e^{(x+\Delta'x+y+\Delta'y)(\theta_1-\theta_2)}-1)\right]. \end{aligned}$$

So $\mathrm{fl}(F'(x,y;d)) = F'(x,y;d) + \delta F'(x,y;d)$ with

$\delta F'(x,y;d)$

$$= \begin{pmatrix} \varepsilon_1+(1+\varepsilon_1)(e^{-\Delta x+\Delta y}-1) & 0 \\ 0 & \varepsilon_2+(1+\varepsilon_2)(e^{-\Delta x-\Delta y}e^{(x+\Delta'x+y+\Delta'y)(\theta_1-\theta_2)}-1) \end{pmatrix}$$

$$\cdot F'(x,y;d) = O(\rho).$$

We can write down an analogous formula for $F''(x,y;d)$.

k	x_6	y_6	ℓ	$\mathrm{cond}(F;e^{10^{-k}})$
0	-0.1000000000000000(01)	0.3597855161523896(-18)	16	$\sqrt{2}$
1	-0.1000000000000000(00)	-0.2376055789464463(-17)	16	$10\sqrt{2}$
2	-0.1000000000000001(-01)	-0.6397150159689099(-17)	15	$10^2\sqrt{2}$
3	-0.0999999999999997(-02)	0.5077502606368951(-17)	15	$10^3\sqrt{2}$
4	-0.0999999999999844(-03)	0.3913464269882279(-17)	13	$10^4\sqrt{2}$
5	-0.0999999999997470(-04)	-0.3905797959965137(-17)	12	$10^5\sqrt{2}$
6	-0.0999999999986935(-05)	0.5633677343553680(-17)	11	$10^6\sqrt{2}$
7	-0.1000000000174599(-06)	-0.1058449777227516(-16)	10	$10^7\sqrt{2}$
8	-0.1000000000015281(-07)	0.4124494865312562(-17)	11	$10^8\sqrt{2}$
9	-0.1000000007452433(-08)	-0.2449359520991520(-17)	9	$10^9\sqrt{2}$
10	-0.0999999914314586(-09)	0.4265833288825851(-17)	8	$10^{10}\sqrt{2}$
11	-0.1000000261210709(-10)	-0.6446772724219823(-17)	7	$10^{11}\sqrt{2}$
12	-0.0999980430668081(-11)	0.3302303528672576(-17)	5	$10^{12}\sqrt{2}$
13	-0.0999761308551817(-12)	0.1322187990417560(-16)	4	$10^{13}\sqrt{2}$
14	-0.1000372750236664(-13)	-0.1182870095748150(-16)	4	$10^{14}\sqrt{2}$
15	-0.0963108239652912(-14)	0.1398012990192197(-17)	2	$10^{15}\sqrt{2}$
16	-0.0868560967896870(-15)	0.3349523961106902(-17)	1	$10^{16}\sqrt{2}$

We remark that the algorithm even behaves considerably well for a condition number of the order of 10^3 or 10^4.

The two linear systems of equations are well-conditioned since the condition number of the linear systems in $x^* = \lim_{i\to\infty} x_i$ is

$$\|F_x'(x^*;d)^{-1}\| \cdot \|F_x'(x^*;d)\| = 2.$$

One can prove that the use of Gaussian elimination with row pivoting for this example satisfies the conditions (d) and (e) of Theorem 2.2. So we can expect to get a reasonable approximation of the solution of $F(x,y;d) = 0$ using the numerically stable iterative method (4); the numerical results illustrate this. Let us at the same time follow the loss of significant digits in the root x^* as the problem becomes worse-conditioned. The calculations are performed in double precision ($t = 56$) on the PDP 11/45 of the University of Antwerp. We will solve the nonlinear system

$F(x, y; d) = 0$ for $d = e^{10^{-k}}$, $k = 0, \ldots, 16$. The root $x^* = (-10^{-k}, 0)$. For each d we give the 6th iteration-step (x_6, y_6) in the procedure (4) starting from $(x_0, y_0) = (2, 2)$, the number l of significant digits in x_6, and the condition number $\operatorname{cond}(F; e^{10^{-k}})$. It is also important to know that the iterative procedure stops at the 6th iteration-step, except for $k = 7$, 13, and 14 where, respectively, $l = 11$, 5, and 3 in the last iteration-step (x_7, y_7). We have used the stop-criterion

$$\max(|x_{i+1} - x_i|, |y_{i+1} - y_i|) \leqslant 10^{-15} \max(|x_{i+1}|, |y_{i+1}|).$$

Department of Mathematics
University of Antwerp, U.I.A.
Universiteitsplein 1
B-2610 Wilrijk, Belgium

1. R. G. Bartle, *The Elements of Real Analysis*, Wiley, New York, 1976.

2. Annie A. M. Cuyt, *Abstract Padé Approximants in Operator Theory*, Lecture Notes in Math., Vol. 765, Springer-Verlag, Berlin and New York, 1979, pp. 61–87.

3. Annie A. M. Cuyt, "On the properties of Abstract Rational (1-point) Approximants," *J. Operator Theory*, v. 5, 1981. (To appear.)

4. A. Cuyt & P. Van der Cruyssen, *Abstract Padé Approximants for the Solution of a System of Nonlinear Equations*, Report 80-17, University of Antwerp, 1980.

5. Louis B. Rall, *Computational Solution of Nonlinear Operator Equations*, Wiley, New York, 1969; reprinted by Krieger, Huntington, New York, 1979.

6. J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, N. J., 1964.

7. H. Woźniakowski, "Numerical stability for solving nonlinear equations," *Numer. Math.*, v. 27, 1977, pp. 373–390.

8. D. Young, *A Survey of Numerical Mathematics*. I, Addison-Wesley, Reading, Mass., 1972.