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

By Annie A. M. Cuyt*

Abstract. Let $F: \mathbb{R}^q \rightarrow \mathbb{R}^q$ and $x^*$ a simple root in $\mathbb{R}^q$ 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 $x_i$, $a_i = -F_{i-1}^{-1} F'_i$ 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 $F''$ evaluated in $(a_i, a_i)$, and componentwise multiplication and division in $\mathbb{R}^q$. 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

(1) $F(x) = 0$

with $F: \mathbb{R}^q \rightarrow \mathbb{R}^q$: $x \rightarrow F(x)$, 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 $d \in \mathbb{R}^p$, called data vector

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

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

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

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

$$||\Delta x|| \leq C_1 \rho ||x||, \quad ||\Delta d|| \leq C_2 \rho ||d||,$$

$$||\Delta F|| \leq C_3 \rho (\Delta F \text{ a } q \times q \text{ matrix}),$$

Received May 27, 1981.

1980 Mathematics Subject Classification. Primary 65G05.

* Aspirant of the Belgian National Fund for Scientific Research (NFWO).
for constants $C_1$, $C_2$, $C_3$ (only depending on the dimensions of the problem) and with $p = 2^t$ 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 $fl(d)$ denote the $t$ digit binary representation of the vector $d$ in floating-point arithmetic

$$\|fl(d) - d\| < Cp\|d\|, \text{ i.e. } fl(d) - d = O(\rho).$$

Since $d$ is represented by $fl(d)$, we solve in fact $F(x; fl(d)) = 0$ 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 $F(x; fl(d)) = 0$ has a root $x^*$ in the neighborhood of $x^*$ and $x^* - x^* = O(\rho)$ if $t$ is sufficiently large; thus,

$$\widetilde{x}^* - x^* = -F'_x(x^*; d)^{-1}F'_d(x^*; d)(fl(d) - d)$$

$$+ \text{higher order terms in } \widetilde{x}^* - x^* \text{ and } fl(d) - d$$

$$= -F'_x(x^*; d)^{-1}F'_d(x^*; d)(fl(d) - d) + O(\rho^2),$$

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

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

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

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

Let us now suppose that $F(x; d) = 0$ is solved by an iterative procedure $\Phi(x, F)$, where $\Phi$ can use several $F_j(x)$, the $j$th Fréchet-derivative of $F$ at $x$ (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 $\widetilde{x}^*$,

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

So

$$\|x_i - x^*\| \leq \|x_i - \widetilde{x}^*\| + \|\widetilde{x}^* - x^*\| \leq K\rho\|\widetilde{x}^*\| + C\rho \cdot \|x^*\| + O(\rho^2)$$

$$< K\rho(\|\widetilde{x}^* - x^*\| + \|x^*\|) + C\rho \cdot \|x^*\| + O(\rho^2)$$

$$< [K\rho + C\rho \cdot \text{cond}(F; d)] \cdot \|x^*\| + O(\rho^2).$$

**Definition 1.2.** An iteration $\Phi$ is called numerically stable if

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

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 $\|x_i - x^*\| < \varepsilon \cdot \|x^*\|$. This is possible if the problem is sufficiently well-conditioned, i.e., $\rho \cdot \text{cond}(F; d) = O(\varepsilon)$. In floating-point arithmetic we have

$$x_{i+1} = \Phi(x_i, F) + \xi_i, \text{ where } \xi_i = fl(\Phi(x_i, F)) - \Phi(x_i, F).$$
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 $\lim_{i \to \infty} ||\xi_i|| < \rho ||x^*||(C \text{cond}(F; d) + K) + O(\rho^2)$.

Proof. We simply verify the definition.

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

2. Abstract Padé Approximants (APA) and Abstract Rational Approximants (ARA) for the Solution of a System of Nonlinear Equations. Let $x_i$ be the $i$th 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

\[ x^* = x_i + a_i + \frac{1}{2} F_i^{-1} F_i'' a_i^2 + O(a_i^3), \]

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].


Theorem 2.1. If
(a) $\text{fl}(F(x_i; d)) = (I + \Delta F_i) F(x_i + \Delta x_i; d + \Delta d_i) = F(x_i) + \delta F_i$, 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) $\text{fl}(F'(x_i; d)) = F'(x_i) + \delta F'_i$, with $\delta F'_i = O(\rho)$,
(c) the computed correction $\text{fl}(a_i)$ is the exact solution of a perturbed linear system
\[ (F'(x_i) + \delta F'_i + E_i) a_i = -F(x_i) - \delta F_i \quad \text{with} \quad E_i = O(\rho), \]

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.

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

where multiplication and division of the vectors in $\mathbb{R}^q$ 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 $q > 1$. 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

\[ \left( \frac{1}{a_i + \frac{1}{2} F_i^{-1} F_i'' a_i^2} \right)^j O(||a_i||^{j-k} \rho^{j-k}) = O(\rho^j). \]
Condition (5) takes care of the fact that the denominator of the correction-term in (4) does not become too small in comparison with \(O(\|a_i\|^{-k}\rho^k)\).

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

\[
\lim_{i \to \infty} \frac{a_i}{a_i + \frac{1}{2} F_i^r F_i^a a_i^2} = 1,
\]

and so \(\exists L \in \mathbb{N} \supseteq \forall i > L: \left| \frac{a_i}{a_i + \frac{1}{2} F_i^r F_i^a a_i^2} \right| < 1 + D\)

(cases \(j = 1, k = 0, l = 0\)) with \(D \in \mathbb{R}_+\), in a convergent process (4): \(\lim_{i \to \infty} \|x^* - x_i\| = 0\), and thus

\[
\lim_{i \to \infty} a_i = 0, \quad \text{i.e. } \exists M \in \mathbb{N} \supseteq \forall i > M: a_i = O(\rho),
\]

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

\[
\lim_{i \to \infty} \frac{a_i^2}{a_i + \frac{1}{2} F_i^r F_i^a a_i^2} = 0, \quad \text{i.e.}
\]

\[
\exists N \in \mathbb{N} \supseteq \forall i > N: \frac{a_i^2}{a_i + \frac{1}{2} F_i^r F_i^a a_i^2} = O(\rho),
\]

and so \(\forall i > \max(N, M): \frac{1}{a_i + \frac{1}{2} F_i^r F_i^a a_i^2} O(\|a_i\|\rho) = O(\rho)\)

(cases \(j = 1, k = 0, l = 1\)).

**Theorem 2.2.** If

(a) \(\text{fl}(F(x_i; d)) = (I + \Delta F_i)F(x_i + \Delta x_i; d + \Delta d_i) = F(x_i) + \delta F_i\) with \(\delta F_i = \Delta F_iF(x_i) + F'_i(x_i) \Delta x_i + F''_i(x_i) \Delta d_i + O(\rho^2)\),

(b) \(\text{fl}(F'(x_i; d)) = F'(x_i) + \delta F'_i \) with \(\delta F'_i = O(\rho)\),

(c) \(\text{fl}(F''(x_i; d)) = F''(x_i) + \delta F''_i \) with \(\delta F''_i = O(\rho)\),

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

\[
(F'(x_i) + \delta F'_i + E_{i1})\text{fl}(a_i) = -F(x_i) - \delta F_i \quad \text{with } E_{i1} = O(\rho),
\]

(e) analogously,

\[
(F'(x_i) + \delta F'_i + E_{i2})\text{fl}(b_i) = (F''(x_i) + \delta F''_i)\text{fl}(a_i)^2
\]

with \(E_{i2} = O(\rho)\) and \(b_i = F_i^{-1} F_i^a a_i^2\),

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

**Proof.** Let \(F'(x_i) + \delta F'_i + E_{i1} = F'(x_i)(I + H_{i1})\), where

\[
H_{i1} = F'(x_i)^{-1}\{\delta F'_i + E_{i1}\} = O(\rho)
\]

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

\[
(I + H_{i1})^{-1} = I - H_{i1} + O(\rho^2).
\]
Thus
\[ \text{(6)} \quad f_l(a_i) = (I - H_{i,1})F_i^{-1}(-F_i - \delta F_i). \]

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

Now
\[
(F_i'' + \delta F_i'')f_l(a_i)^2 = (F_i'' + \delta F_i'') \left[ (I - H_{i,1})F_i^{-1}(-F_i - \delta F_i) \right]^2
\]
\[
= (F_i'' + \delta F_i'')a_i^2 + 2(F_i'' + \delta F_i'')(F_i''F_i' - H_{i,1}F_i'^{-1}F_i) + O(\rho^2)
\]
\[
= (F_i'' + \delta F_i'')a_i^2 - 2F_i''(a_i, F_i''\delta F_i - H_{i,1}F_i'^{-1}F_i) + O(\rho^2).
\]

Thus
\[ f_l(b_i) = F_i^{-1}(F_i'' + \delta F_i'')a_i^2 - 2F_i'^{-1}F_i''(a_i, F_i''\delta F_i - H_{i,1}F_i'^{-1}F_i)
\]
\[ \quad - H_{i,2}F_i^{-1}F_i''a_i^2 + O(\rho^2). \]

A computed approximation \( x_{i+1} \) satisfies
\[
x_{i+1} = (I + \delta I_{i,1}) \left[ x_i + (I + \delta I_{i,2}) \frac{f_l(a_i)^2}{f_l(a_i) + \frac{1}{2}f_l(b_i)} \right],
\]
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
\[
x_{i+1} = (I + \delta I_{i,1}) \left[ x_i + (I + \delta I_{i,2}) \frac{a_i^2 - 2 a_i \cdot (F_i'^{-1}\delta F_i + H_{i,1}a_i) + O(\rho^2)}{a_i + \frac{1}{2}b_i - \delta a_i + O(\rho^2)} \right],
\]
where
\[
\delta a_i = F_i'^{-1}\delta F_i + H_{i,1}a_i - \frac{1}{2}F_i'^{-1}\delta F_i''a_i^2
\]
\[ + \frac{1}{2}H_{i,2}F_i'^{-1}F_i''a_i^2 + F_i'^{-1}F_i''(a_i, F_i'^{-1}\delta F_i - H_{i,1}F_i'^{-1}F_i). \]

Using (6), we find
\[ f_l(a_i) - a_i + H_{i,1}a_i - H_{i,1}F_i'^{-1}\delta F_i = -F_i'^{-1}\delta F_i, \]
and thus, for positive constants \( D_1 \) and \( D_2 \),
\[ \|F_i'^{-1}\delta F_i\| < D_2\rho\|a_i\| \quad \text{since } \|f_l(a_i) - a_i\| < D_1\rho\|a_i\| \]
and
\[ \|F_i'^{-1}\| \cdot \|F_i\| < \|F_i'^{-1}\| \cdot \|F_i\| \cdot \|a_i\|. \]

Thus
\[
x_{i+1} = (I + \delta I_{i,1}) \left[ x_i + \frac{a_i^2 - 2 a_i \cdot (F_i'^{-1}\delta F_i + H_{i,1}a_i) + \delta I_{i,2}a_i^2 + O(\rho^2\|a_i\|^2)}{a_i + \frac{1}{2}b_i - \delta a_i + O(\rho^2)} \right],
\]
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
\[
x_{i+1} = (I + \delta I_{i,1}) \left[ x_i + \frac{a_i^2 - 2 a_i \cdot (F_i'^{-1}\delta F_i + H_{i,1}a_i) + \delta I_{i,2}a_i^2 + O(\rho^2\|a_i\|^2)}{a_i + \frac{1}{2}b_i} \right] c_i,
\]
with
\[ c_i = 1 + \frac{1}{a_i + \frac{1}{2} b_i} (\delta a_i + O(\rho^2)) + \left( \frac{1}{a_i + \frac{1}{2} b_i} \right)^2 O(||a_i||^2 - k \rho^k, k = 0, 1, 2) \]
since \( \delta a_i = O(\rho||a_i||) \); in \( c_i \) we have used the notation \( \mathbf{1} \) for the unit vector \((1, \ldots, 1)\).

Using (5), we conclude
\[ \left( \frac{1}{a_i + \frac{1}{2} b_i} \right)^2 O(||a_i||^2 - k \rho^k, k = 0, 1, 2) = O(\rho^2). \]

For \( \xi_i = x_{i+1} - \Phi(x_i, F) \), we have
\[ \xi_i = \delta I_{i,1} x_i + \frac{a_i^2}{a_i + \frac{1}{2} b_i} (c_i - 1) \]
\[ + \frac{-2a_i (F_i^{-1} \delta F_i + H_{i,x_i} a_i)}{a_i + \frac{1}{2} b_i} + O(\rho^2||a_i||^2) \cdot c_i \]
\[ + \delta I_{i,1} \frac{a_i^2}{a_i + \frac{1}{2} b_i} \cdot c_i + O(\rho^2). \]

So
\[ \xi_i = \delta I_{i,1} x_i + \left( \frac{1}{a_i + \frac{1}{2} b_i} \right)^2 O(\rho||a_i||^3, \rho^2||a_i||^2) + \frac{1}{a_i + \frac{1}{2} b_i} O(\rho^2||a_i||^2) \]
\[ + \frac{1}{a_i + \frac{1}{2} b_i} (-2a_i F_i^{-1} \delta F_i + O(\rho||a_i||^2, \rho^2||a_i||^2)) \cdot (1 + O(\rho)) \]
\[ + O(\rho^2). \]

Thus
\[ ||\xi_i|| < k_1 \rho||x_i|| + k_2 \rho||a_i|| + \frac{-2a_i}{a_i + \frac{1}{2} b_i} F_i^{-1} \delta F_i \]
\[ + O(\rho^2), \]
and since
\[ -\frac{-2a_i}{a_i + \frac{1}{2} b_i} F_i^{-1} \delta F_i = -\frac{-2a_i}{a_i + \frac{1}{2} b_i} F_i^{-1} (\Delta F_i F(x_i) + F_i F(x_i) + F_i F_d \Delta d_i + O(\rho^2)) \]
\[ = \frac{1}{a_i + \frac{1}{2} b_i} O(\rho||a_i||) F(x_i) - \frac{2a_i}{a_i + \frac{1}{2} b_i} \Delta x_i \]
\[ - \frac{2a_i}{a_i + \frac{1}{2} b_i} F_i^{-1} F_d \Delta d_i + \frac{1}{a_i + \frac{1}{2} b_i} O(\rho^2||a_i||), \]
we find that
\[ \lim_{i \to \infty} ||\xi_i|| < \rho||x^*|| (K + C \text{ cond}(F; d)) + O(\rho^2) \]
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 \( a_i F_i^{-1} F_d \Delta d_i = O(\rho||a_i||) \).
3. Numerical Example. Consider the following operator:

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

The operator $F$ has a simple root $x^* = (-\frac{1}{2} \ln(d_1d_2), \frac{1}{2} \ln(d_1/d_2))$. Clearly

$$d = (d_1, d_2)$$

is the data vector. Now

$$fl(F(x, y; d)) = \begin{bmatrix} (1 + \epsilon_1)e^{(x+\Delta x+\Delta y)(1+\kappa_1)} - (d_1 + \Delta_1 d) \\ (1 + \epsilon_2)e^{(x-\Delta x-\Delta y)(1+\kappa_2)} - (d_2 + \Delta_2 d) \end{bmatrix}(1 + \kappa_1),$$

where $fl(x) = x + \Delta x$, $fl(y) = y + \Delta y$, $fl(d_1) = d_1 + \Delta_1 d$, $fl(d_2) = d_2 + \Delta_2 d$, $\theta_1$ is caused by $-fl(x) + fl(y)$, $\theta_2$ is caused by $-fl(x) - fl(y)$, $\epsilon_i$ are caused by the exponential evaluations ($i = 1, 2$), $\kappa_i$ are caused by the subtraction of $fl(d_i)$ ($i = 1, 2$).

One can rewrite $fl(F(x, y; d)) = (I + AF)F(x + Ax, y + Ay; d + Ad)$ with

$$\Delta x = x\theta_1 + \Delta x(1 + \kappa_1), \quad \Delta y = y\theta_1 + \Delta y(1 + \kappa_1), \quad \Delta d = (\Delta_1 d, \Delta_2 d),$$

$$\Delta_1 d = \frac{\Delta_1 d - \epsilon_1 d_1}{1 + \epsilon_1},$$

$$\Delta_2 d = \frac{\Delta_2 d - \epsilon_2 d_2}{1 + \epsilon_2} + \frac{d_2 + \Delta_2 d}{1 + \epsilon_2} \left(e^{x+\Delta x+\Delta y}(\theta_1+\epsilon_i) - 1\right),$$

$$\Delta F = \begin{pmatrix} (1 + \epsilon_1)(1 + \kappa_1) - 1 & 0 \\ 0 & (1 + \epsilon_2)(1 + \kappa_2) e^{x+\Delta x+\Delta y}(\theta_1+\epsilon_i) - 1 \end{pmatrix}.$$  

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

$$F'(x^*)^{-1} = \begin{pmatrix} -d_2 & -d_1 \\ d_2 & -d_1 \end{pmatrix} \quad \text{and} \quad \Delta F' = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$  

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

$$\|F'(x^*)^{-1}\| \cdot \|F(d, d)\| \quad \text{with} \quad \|F(d, d)\|.$$

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 $\text{cond}(F; d) < 1$, i.e.,

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

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

Now

$$fl(F'(x, y; d)) = fl\begin{pmatrix} -e^{-x+y} & e^{-x+y} \\ -e^{-x-y} & -e^{-x-y} \end{pmatrix}.$$
where
\[ f(e^{-x+y}) = (1 + \varepsilon_1) e^{-x-y+\varepsilon_1} = (1 + \varepsilon_1) e^{-x+y} e^{-\Delta x + \Delta y} \]
\[ = e^{-x+y} \left[ 1 + \varepsilon_1 + (1 + \varepsilon_1)(e^{-\Delta x + \Delta y} - 1) \right] \]
\[ f(e^{-x-y}) = (1 + \varepsilon_2) e^{-x+y-\varepsilon_2} = (1 + \varepsilon_2) e^{-x-y} e^{-\Delta x - \Delta y} e^{(x+y+\Delta x + \Delta y)(\varepsilon_1 - \varepsilon_2)} \]
\[ = e^{-x-y} \left[ 1 + \varepsilon_2 + (1 + \varepsilon_2)(e^{-\Delta x - \Delta y} e^{(x+y+\Delta x + \Delta y)(\varepsilon_1 - \varepsilon_2)} - 1) \right] \]

So \( f(F'(x, y; d)) = F'(x, y; d) + \delta F'(x, y; d) \) with
\[
\delta F'(x, y; d) = \begin{pmatrix}
\varepsilon_1 + (1 + \varepsilon_1)(e^{-\Delta x + \Delta y} - 1) \\
0
\end{pmatrix}
\begin{pmatrix}
0 \\
\varepsilon_2 + (1 + \varepsilon_2)(e^{-\Delta x - \Delta y} e^{(x+y+\Delta x + \Delta y)(\varepsilon_1 - \varepsilon_2)} - 1) \\
\end{pmatrix}
\cdot F'(x, y; d) = O(\rho).

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

<table>
<thead>
<tr>
<th>( k )</th>
<th>( x_6 )</th>
<th>( y_6 )</th>
<th>( t )</th>
<th>( \text{cond}(F_2 e^{10^{-k}}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.10</td>
<td>0.35</td>
<td>16</td>
<td>\sqrt{2}</td>
</tr>
<tr>
<td>1</td>
<td>-0.10</td>
<td>-0.23</td>
<td>16</td>
<td>10^{2/2}</td>
</tr>
<tr>
<td>2</td>
<td>-0.10</td>
<td>-0.62</td>
<td>15</td>
<td>10^{3/2}</td>
</tr>
<tr>
<td>3</td>
<td>-0.10</td>
<td>0.50</td>
<td>15</td>
<td>10^{4/2}</td>
</tr>
<tr>
<td>4</td>
<td>-0.10</td>
<td>0.39</td>
<td>13</td>
<td>10^{5/2}</td>
</tr>
<tr>
<td>5</td>
<td>-0.10</td>
<td>-0.39</td>
<td>12</td>
<td>10^{6/2}</td>
</tr>
<tr>
<td>6</td>
<td>-0.10</td>
<td>0.56</td>
<td>11</td>
<td>10^{7/2}</td>
</tr>
<tr>
<td>7</td>
<td>-0.10</td>
<td>-0.10</td>
<td>10</td>
<td>10^{8/2}</td>
</tr>
<tr>
<td>8</td>
<td>-0.10</td>
<td>0.41</td>
<td>11</td>
<td>10^{9/2}</td>
</tr>
<tr>
<td>9</td>
<td>-0.10</td>
<td>-0.24</td>
<td>9</td>
<td>10^{10/2}</td>
</tr>
<tr>
<td>10</td>
<td>-0.10</td>
<td>0.42</td>
<td>8</td>
<td>10^{11/2}</td>
</tr>
<tr>
<td>11</td>
<td>-0.10</td>
<td>-0.64</td>
<td>7</td>
<td>10^{12/2}</td>
</tr>
<tr>
<td>12</td>
<td>-0.10</td>
<td>0.33</td>
<td>5</td>
<td>10^{13/2}</td>
</tr>
<tr>
<td>13</td>
<td>-0.10</td>
<td>0.13</td>
<td>4</td>
<td>10^{14/2}</td>
</tr>
<tr>
<td>14</td>
<td>-0.10</td>
<td>-0.11</td>
<td>2</td>
<td>10^{15/2}</td>
</tr>
<tr>
<td>15</td>
<td>-0.09</td>
<td>0.13</td>
<td>1</td>
<td>10^{16/2}</td>
</tr>
</tbody>
</table>

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
\[
\left\| F'_x(x^*; d)^{-1} \right\| \cdot \left\| F'_x(x^*; d) \right\| = 2.
\]

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 $\text{cond}(F; e^{10^{-k}})$. It is also important to know that the iterative procedure stops at the 6th iteration-step, except for $k = 7, 13, 14$ where, respectively, $l = 11, 5, 3$ in the last iteration-step $(x_7, y_7)$. We have used the stop-criterion

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

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