МАТЕМАТИЧКИ ВЕСНИК
57 (2005), 27–33

UDK 519.633.6
оригинални научни рад
research paper

NUMERICAL STABILITY OF A CLASS (OF SYSTEMS) OF NONLINEAR EQUATIONS

Zlatko Udovičić

Abstract. In this article we consider stability of nonlinear equations which have the following form:

$$Ax + F(x) = b, \tag{1}$$

where F is any function, A is a linear operator, b is given and x is an unknown vector. We give (under some assumptions about function F and operator A) a generalization of inequality:

$$\frac{\|X_1 - X_2\|}{\|X_1\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|b_1 - b_2\|}{\|b_1\|} \tag{2}$$

(equation (2) estimates the relative error of the solution when the linear equation $Ax = b_1$ becomes the equation $Ax = b_2$) and a generalization of inequality:

$$\frac{\|X_1 - X_2\|}{\|X_1\|} \leq \left\|A_1^{-1}\right\| \|A_1\| \left(\frac{\|b_1 - b_2\|}{\|b_1\|} + \|A_1\| \left\|A_2^{-1}\right\| \frac{\|b_2\|}{\|b_1\|} \cdot \frac{\|A_1 - A_2\|}{\|A_1\|} \right) \tag{3}$$

(equation (3) estimates the relative error of the solution when the linear equation $A_1 x = b_1$ becomes the equation $A_2 x = b_2$).

1. Basic results

Teorem 1. *Let V be a normed space, let the linear operator $A\colon V \to V$ be invertible and bounded, let the inverse operator of the operator A be also bounded, let $b_1, b_2 \in V$ and let the functions $F_1, F_2\colon V \to V$ and the set $S \subseteq V$ have the following properties:*

1. *the function F_1 is Lipschitz on S, i.e.,*

$$(\exists L > 0)\,(\forall x_1, x_2 \in S)\, \|F_1(x_1) - F_1(x_2)\| \leq L \|x_1 - x_2\|,$$

and the constant L is such that the inequality $1 - L\left\|A^{-1}\right\| > 0$, holds;

2. $(\exists M > 0)\,(\forall x \in S)\, \|F_1(x)\| \leq M\|x\|$; *and*
3. $(\exists \varepsilon \geq 0)\,(\forall x \in S)\, \|F_1(x) - F_2(x)\| \leq \varepsilon$.

AMS Subject Classification: 65J15
Keywords and phrases: Numerical stability, nonlinear equations.

If $X_1 \in S$ is a solution of the equation $Ax + F_1(x) = b_1$ and $X_2 \in S$ is a solution of the equation $Ax + F_2(x) = b_2$, then the following inequality holds:

$$\frac{\|X_1 - X_2\|}{\|X_1\|} \leq \frac{\left\|A^{-1}\right\| (\|A\| + M)}{1 - L\left\|A^{-1}\right\|} \left(\frac{\|b_1 - b_2\|}{\|b_1\|} + \frac{\varepsilon}{\|b_1\|} \right)$$

Proof. Since $AX_1 + F_1(X_1) = b_1$, we have

$$\|b_1\| = \|AX_1 + F_1(X_1)\| \leq \|AX_1\| + \|F_1(X_1)\| \leq (\|A\| + M)\|X_1\|$$

and we can conclude that

$$\frac{1}{\|X_1\|} \leq \frac{(\|A\| + M)}{\|b_1\|}. \tag{4}$$

On the other hand, from $X_1 - X_2 = A^{-1}((b_1 - b_2) - (F_1(X_1) - F_2(X_2)))$ it follows that

$$\begin{aligned} \|X_1 - X_2\| &\leq \left\|A^{-1}\right\| (\|b_1 - b_2\| + \|F_1(X_1) - F_1(X_2)\| + \|F_1(X_2) - F_2(X_2)\|) \\ &\leq \left\|A^{-1}\right\| (\|b_1 - b_2\| + L\|X_1 - X_2\| + \varepsilon), \end{aligned}$$

and that

$$\|X_1 - X_2\| \leq \frac{\left\|A^{-1}\right\| (\|b_1 - b_2\| + \varepsilon)}{1 - L\left\|A^{-1}\right\|}. \tag{5}$$

Finally, from (4) and (5) we have

$$\frac{\|X_1 - X_2\|}{\|X_1\|} \leq \frac{\left\|A^{-1}\right\| (\|A\| + M)}{1 - L\left\|A^{-1}\right\|} \left(\frac{\|b_1 - b_2\|}{\|b_1\|} + \frac{\varepsilon}{\|b_1\|} \right)$$

which proves the theorem. ∎

If $F_1 \equiv 0$ and $F_2 \equiv 0$ (in this case we have $L = M = \varepsilon = 0$), then the proved inequality becomes (2).

THEOREM 2. *Let V be a normed space, let the linear operators $A_1, A_2 \colon V \to V$ be invertible and bounded, let their inverse operators be also bounded, let $b_1, b_2 \in V$ and let the function $F \colon V \to V$ and the set $S \subseteq V$ have the following properties:*

1. *the function F is Lipschitz on S, i.e.,*

   $$(\exists L > 0)(\forall x_1, x_2 \in S)\, \|F(x_1) - F(x_2)\| \leq L\|x_1 - x_2\|,$$

   and the constant L is such that the inequality $1 - L\left\|A_1^{-1}\right\| > 0$ holds;
2. $(\exists M > 0)(\forall x \in S)\, \|F(x)\| \leq M\|x\|$ *;*
3. *the function F is bounded on the set S, i.e.,*

   $$(\exists B \geq 0)(\forall x \in S)\, \|F(x)\| \leq B.$$

If $X_1 \in S$ is a solution of the equation $A_1x + F(x) = b_1$ and $X_2 \in S$ is a solution of the equation $A_2x + F(x) = b_2$, then the following inequality holds:

$$\begin{aligned} \frac{\|X_1 - X_2\|}{\|X_1\|} \leq \frac{\left\|A_1^{-1}\right\| (\|A_1\| + M)}{1 - L\left\|A_1^{-1}\right\|} \Bigg(&\frac{\|b_1 - b_2\|}{\|b_1\|} + \|A_1\| \left\|A_2^{-1}\right\| \times \\ &\times \frac{\|b_2\|}{\|b_1\|} \cdot \frac{\|A_1 - A_2\|}{\|A_1\|} + \frac{B\left\|I - A_1 \cdot A_2^{-1}\right\|}{\|b_1\|} \Bigg). \end{aligned}$$

Proof. Since $X_2 = A_2^{-1} \cdot (b_2 - F(X_2))$, we have

$$\begin{aligned} A_1 X_2 &= A_1 X_2 + b_2 - A_2 X_2 - F(X_2) \\ &= (A_1 - A_2) X_2 + b_2 - F(X_2) \\ &= (A_1 - A_2) A_2^{-1} (b_2 - F(X_2)) + b_2 - F(X_2) \\ &= (A_1 - A_2) A_2^{-1} b_2 - (A_1 - A_2) A_2^{-1} F(X_2) + b_2 - F(X_2) \\ &= (A_1 - A_2) A_2^{-1} b_2 + b_2 - A_1 A_2^{-1} F(X_2), \end{aligned}$$

and we can apply the previous theorem to the equations

$$A_1 x + F(x) = b_1$$

and

$$A_1 x + A_1 A_2^{-1} F(x) = (A_1 - A_2) A_2^{-1} b_2 + b_2.$$

Condition 3. of the theorem is satisfied since for every $x \in S$ the inequality

$$\left\| F(x) - A_1 A_2^{-1} F(x) \right\| \le \| F(x) \| \left\| I - A_1 A_2^{-1} \right\| \le B \left\| I - A_1 A_2^{-1} \right\|$$

holds. So,

$$\begin{aligned} \frac{\|X_1 - X_2\|}{\|X_1\|} &\le \frac{\left\|A_1^{-1}\right\| (\|A_1\| + M)}{1 - L \left\|A_1^{-1}\right\|} \left(\frac{\left\| b_1 - b_2 - (A_1 - A_2) A_2^{-1} b_2 \right\|}{\|b_1\|} + \right. \\ &\qquad \left. + \frac{B \left\| I - A_1 A_2^{-1} \right\|}{\|b_1\|} \right) \\ &\le \frac{\left\|A_1^{-1}\right\| (\|A_1\| + M)}{1 - L \left\|A_1^{-1}\right\|} \left(\frac{\|b_1 - b_2\|}{\|b_1\|} + \|A_1\| \left\|A_2^{-1}\right\| \times \right. \\ &\qquad \left. \times \frac{\|b_2\|}{\|b_1\|} \cdot \frac{\|A_1 - A_2\|}{\|A_1\|} + \frac{B \left\| I - A_1 \cdot A_2^{-1} \right\|}{\|b_1\|} \right). \end{aligned}$$

The theorem has been proved. ■

If $F \equiv 0$ (in this case we have $L = M = B = 0$), then the inequality just proved becomes (3).

From the theorems just proved we can conclude that relatively small changes (of operator A, function F or vector b) in the equation (1) may cause relatively big changes in the solution if the number

$$\frac{\left\|A^{-1}\right\| (\|A\| + M)}{1 - L \left\|A^{-1}\right\|} \tag{6}$$

is big enough, so we can take this number as a measure of stability of equation (1). It is obvious that the equation (1) gets more badly conditioned as the number (6) increases. Since the inequality $\|A\|\|A^{-1}\| > 1$ always holds, the number (6) is greater than one whenever inequality $1 - L\|A^{-1}\| > 0$ holds.

2. A note

If the normed space X is complete and the subset $S \subseteq X$ is closed, if the function F satisfies the condition 1. of Theorem 1 (Theorem 2) and if $\varphi(S) \subseteq S$ where $\varphi(x) = A^{-1}(b - F(x))$, then the array generated by the recursive formula

$$x_{n+1} = A^{-1}(b - F(x_n)), n \in \mathbb{N} \tag{7}$$

converges to the unique solution of the equation (1) for every $x_0 \in S$.

Indeed, the function φ is a contraction since for every $x, y \in S$ we have

$$\|\varphi(x) - \varphi(y)\| \le \left\|A^{-1}\right\| \|b - F(x) - b + F(y)\| \le L \left\|A^{-1}\right\| \|x - y\|,$$

while from the condition 1. of Theorem 1 (Theorem 2) we have that $L\left\|A^{-1}\right\| < 1$, and therefore in accordance with Banach fixed point theorem, the array defined by formula (7) will converge to the unique solution of the equation (1).

3. Examples

The first example will give (under certain assumptions) a sufficient condition for stability of polynomial with real coefficients. We thoroughly considered polynomials of the third degree.

Example 1. Let a polynomial with real coefficients $P(x) = ax^3 + bx^2 + cx + d$, $(a, c \neq 0)$ have at least one zero in the segment $[\alpha, \beta]$. Furthermore, let $F(x) = ax^3 + bx^2$ and let $\Lambda = \max\{|\alpha|, |\beta|\}$. Then we have that $(\forall x \in [\alpha, \beta])\, |F(x)| \le \left(|a|\Lambda^2 + |b|\Lambda\right)|x|$ and $\max_{x\in[\alpha,\beta]} |F'(x)| = \max\left\{|F'(\alpha)|, |F'(\beta)|, \left|F'\left(-\frac{b}{3a}\right)\right|\right\}$ and in Theorems 1 and 2 we can put that

$$M = |a|\Lambda^2 + |b|\Lambda,$$

and that

$$L = \max\left\{|F'(\alpha)|, |F'(\beta)|, \left|F'\left(-\frac{b}{3a}\right)\right|\right\}.$$

If the condition $1 - \frac{L}{|c|} > 0 \iff L < |c|$ is satisfied then, in accordance with Theorems 1 and 2 we can say that if the number $\frac{|c|+M}{|c|-L} = \frac{1+M/|c|}{1-L/|c|}$ (which is always greater than one) is close enough to one, then relatively small changes in coefficients of the polynomial P will not cause relatively great changes in roots of the polynomial. So, if linear term in polynomial P is more dominant ($|c| \gg M$ and $|c| \gg L$), the polynomial P is better conditioned.

We can do the same thing with polynomial of the fourth degree $P(x) = ax^4 + bx^3 + cx^2 + dx + e$, $(a, d \neq 0)$ and conclude that the number $\frac{|d|+M}{|d|-L}$ (the numbers M and L have the same meaning) can be used as a measure of stability of the polynomial P. So, the polynomial P is in this case also better conditioned if the number $\frac{|d|+M}{|d|-L}$ is closer to one. Of course, we can use the same technics for the polynomials of higher degrees, but in that case the problem of effective finding of number L is much more complex.

Example 2. Let V be a normed space, let $d \in V$ be a fixed vector, and let the function $F\colon V \to V$ be defined by

$$(\forall x \in V)\, F(x) = \|x\|\, d.$$

We shall consider the relative error of solution when the equation $A_1 x + F(x) = b$ becomes the equation $A_2 x + F(x) = b$. Since for every $x, x_1, x_2 \in V$ inequality

$$\|F(x_1) - F(x_2)\| \le \|x_1 - x_2\|\, \|d\|$$

and equality

$$\|F(x)\| = \|x\|\, \|d\|\,,$$

hold, we can put $L = M = \|d\|$. So, if the condition $1 - \|d\| \left\|A_1^{-1}\right\| > 0$, is satisfied we can take the number $\dfrac{\left\|A_1^{-1}\right\| (\|A_1\| + \|d\|)}{1 - \|d\| \left\|A_1^{-1}\right\|}$ as a measure of stability for the considered equation.

The following example is a numerical realization of Example 2.

Example 3. The solution of the system

$$\begin{aligned} \max\{x, y\} + 2.01x - 1000y &= 1000 \\ \max\{x, y\} - 0.01x - 1000y &= -1000 \end{aligned}$$

is $X_1 = \begin{pmatrix} 990.099 \\ 1.98020 \end{pmatrix}$, while the solution of the system

$$\begin{aligned} \max\{x, y\} + 2.02x - 1000y &= 1000 \\ \max\{x, y\} - 0.01x - 1000y &= -1000, \end{aligned}$$

is the vector $X_2 = \begin{pmatrix} 985.222 \\ 1.97537 \end{pmatrix}$.

Stability of the considered system can be estimated by using the previous example ($V = \mathbb{R}^2$, $d = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and the norm is the uniform norm of the space $\mathbb{R}^2$). The relative error of the matrix $A_1 = \begin{bmatrix} 2.01 & -1000 \\ -0.01 & -1000 \end{bmatrix}$ when this matrix becomes the matrix $A_2 = \begin{bmatrix} 2.02 & -1000 \\ -0.01 & -1000 \end{bmatrix}$ is $\dfrac{\|A_1 - A_2\|_\infty}{\|A_1\|_\infty} \approx 10^{-5}$ ($10^{-3}\%$), while the relative error of the solution when the first system becomes the second one is $\dfrac{\|X_1 - X_2\|_\infty}{\|X_1\|_\infty} \approx 0.5 \cdot 10^{-2}$ (0.5%). So, the relative error of the solution is approximately 500 times bigger then the relative error of the matrix A. According to the proved theorems our system is badly conditioned since $\dfrac{\left\|A_1^{-1}\right\|_\infty (\|A_1\|_\infty + \|d\|_\infty)}{1 - \|d\|_\infty \left\|A_1^{-1}\right\|_\infty} = 100301.$

It should be noted that the influence of nonlinear term in this example is irrelevant. The relative error of solution, when linear system $A_1x = b = \begin{pmatrix} 1000 \\ -1000 \end{pmatrix}$ becomes the system $A_2x = b$ is approximately 0.5%, too.

We would like to point out that this system may also be solved by using the Banach fixed-point theorem (see Section 2).

The first one of the following examples has a theoretical character, while the second one is its numerical realization.

Example 4. Let V be a normed space, let $d \in V$ and $r > 0$ be a fixed vector and a real number, let $S = \{x \in V \mid \|x\| \leq r\}$ and let the function $F\colon V \to V$ be defined by

$$(\forall x \in V)\, F(x) = \|x\|^2 d.$$

We shall estimate the relative error of the solution when equation $A_1x + F(x) = b$ becomes equation $A_2x + F(x) = b$. Since for every $x, x_1, x_2 \in S$ inequalities

$$\begin{aligned} \|F(x_1) - F(x_2)\| &= \left\| \|x_1\|^2 d - \|x_2\|^2 d \right\| \\ &= (\|x_1\| + \|x_2\|) \cdot \left| \|x_1\| - \|x_2\| \right| \cdot \|d\| \\ &\leq 2r \cdot \|d\| \cdot \|x_1 - x_2\| \end{aligned}$$

and

$$\|F(x)\| = \|x\|^2 \|d\| \leq r \|d\| \|x\|,$$

hold, we can put $M = r\|d\|$ and $L = 2r\|d\|$. So, if the condition $1 - 2r\|d\| \left\|A_1^{-1}\right\| > 0$ is satisfied then the number $\dfrac{\left\|A_1^{-1}\right\| (\|A_1\| + r\|d\|)}{1 - 2r\|d\| \left\|A_1^{-1}\right\|}$ can be taken as a measure of stability of the considered equation.

Example 5. The solution of the system

$$\begin{aligned} x^2 + y^2 + 750x + 50y &= -1 \\ x^2 + y^2 + 2x - 3y &= -1 \end{aligned}$$

which belongs to the set $S = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1 \right\}$ is $X_1 = \begin{pmatrix} -0.0254856 \\ 0.359684 \end{pmatrix}$, while the solution of the system

$$\begin{aligned} x^2 + y^2 + 750x + 50y &= -1 \\ x^2 + y^2 + 2x - 2y &= -1 \end{aligned}$$

which belongs to the set $S = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1 \right\}$ is the vector $X_2 = \begin{pmatrix} -0.0481967 \\ 0.693291 \end{pmatrix}$.

Stability of the considered system can be estimated by using Example 4 ($V = \mathbb{R}^2$, $S = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1 \right\}$, $d = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, while the norm is the Euclidean norm of the space $\mathbb{R}^2$). Relative error of the matrix $A_1 = \begin{bmatrix} 750 & 50 \\ 2 & -3 \end{bmatrix}$ when this matrix becomes the matrix $A_2 = \begin{bmatrix} 750 & 50 \\ 2 & -2 \end{bmatrix}$ is $\dfrac{\|A_1 - A_2\|_2}{\|A_1\|_2} \approx 0.13 \cdot 10^{-2}$ (0.13%), while the relative error of the solution when the first system becomes the second one is $\dfrac{\|X_1 - X_2\|_2}{\|X_1\|_2} \approx 0.93$ (93%). So, the relative error of the solution is approximately 700 times bigger than the relative error of matrix A. According to the proved theorems the system is badly conditioned since $\dfrac{\left\|A_1^{-1}\right\|_2 (\|A_1\|_2 + \|d\|_2)}{1 - 2\,\|d\|_2 \left\|A_1^{-1}\right\|_2} = 2527.$

Contrary to Example 3, the influence of nonlinear term is important now. In this example, the relative error of solution when linear system $A_1 x = b = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ becomes the system $A_2 x = b$ is approximately 47% .

REFERENCES

[**1**] Bahvalov, N. S., *Numerical Methods*, Science, Moscow, 1973.

[**2**] Edwards, R. E., *Functional Analysis*, Holt, Rinehart and Winston, Inc., New York, 1965.

(received 12.05.2004, in revised form 14.02.2005)

Faculty of Sciences, Department of Mathematics, Zmaja od Bosne 35, 71000 Sarajevo, Bosnia and Herzegovina

E-mail: `zzlatko@pmf.unsa.ba`