Bull. Nov. Comp. Center, Num. Anal., 5 (1994), 55–65

Boundedness conditions and the direct analysis of computation errors of a recurrent relation

A.N. Ostilovsky

In this paper the recurrent relation generating the chain fraction is considered. We find different non-weakened sufficient conditions of its boundedness and small growth of computation errors, caused by the inaccuracy of arithmetical operations implementation. These conditions as well as error estimates do not contain any indefinite constants and are expressed in the form of exact inequalities through machine ε.

1. Introduction

In this paper the recurrent relation

$$c_1 = 0, \quad \beta_0 = 0, \quad \beta_k = \frac{-a_k}{b_k + c_k\beta_{k-1}}, \quad k = 1, 2, \ldots, \tag{1.1}$$

is considered. This relation appears, for example, in the solution of the system of linear algebraic equations with the matrix

$$A = \text{tridiag}(c_k, b_k, a_k) \tag{1.2}$$

by the sweep method.

It is well-known that under conditions of the diagonal dominance, i.e.,

$$\left|\frac{a_k}{b_k}\right| + \left|\frac{c_k}{b_k}\right| \le r \le 1, \quad k = 1, 2, \ldots, \tag{1.3}$$

and when at least one of these inequalities is rigorous, the following estimate takes place:

$$|\beta_k| \le r, \quad k = 1, 2, \ldots . \tag{1.4}$$

In the present paper it is shown that for $r < 1$ condition (1.3) can be weakened. In Section 2 the other non-weakened conditions are found, which are sufficient for (1.4). For this case $r \le 1$ is not assumed. In Section 3 sufficient conditions for bondedness of $|\beta_k|$ are found. Then the estimate

of sup $|\beta_k|$ is obtained. In Section 3 we will note, particularly, the mutual disposition of the domains $\nu(r)$ for different r.

The direct analysis of errors in computer realization (1.1) is made on the exact model of machine arithmetic unlike [1]. All errors are expressed through ε_M, namely, the machine relative error of unity. The estimations for the total relative errors are given not exactly to within small highest orders, but as true inequalities. These estimations do not contain the indefinite constants. The conditional noumber is not present in them in the explict form as well. The restrictions on c_k, b_k, a_k are found, for which the growth of the relative computation error β_k is inessential (the details are in the text). Under these restrictions the smoothness of coefficients is not required. For the analysis of errors the theoretical-group methods are used.

2. Boundedness of β_k. The direct problem

Let us assume that all $b_k = 1$ in (1.1), i.e., we will consider

$$c_1 = 0, \quad \beta_0 = 0, \quad \beta_k = \frac{-a_k}{1 + c_k\beta_{k-1}}, \quad k = 1, 2, \ldots . \tag{2.1}$$

Let $r > 0$ be set. Our aim is to indicate the non-recurrent condition on c_k, a_k in (2.1) sufficient for

$$|\beta_k| \le r, \quad k = 1, 2, \ldots . \tag{2.2}$$

It is well-known that one of such conditions for $r < 1$ is a condition of diagonal dominance

$$(c_k, a_k) \in \mu(r) = \{(c, a): \ |c| + |a| \le r\}, \quad k = 1, 2, \ldots . \tag{2.3}$$

Let us weaken (2.3) and offer a number of other conditions providing (2.2).

On the coordinate plane cOa we will represent the domains $\omega_{ij}(r)$, $\nu(r)$, which are given by the systems of inequalities (see Figure 1)

$$\omega_{ij}(r) = \begin{cases} (-1)^i r^2 c + (-1)^j a + r \ge 0, \\ 0 \le (-1)^{j-1} a \le r, \\ c \ne (-1)^{i-1} \dfrac{1}{r}, \end{cases} \qquad i, j = 1, 2, \tag{2.4}$$

$$\nu(r) = \begin{cases} r^2|c| + |a| \le r, \\ |c| \ne \dfrac{1}{r}. \end{cases} \tag{2.5}$$

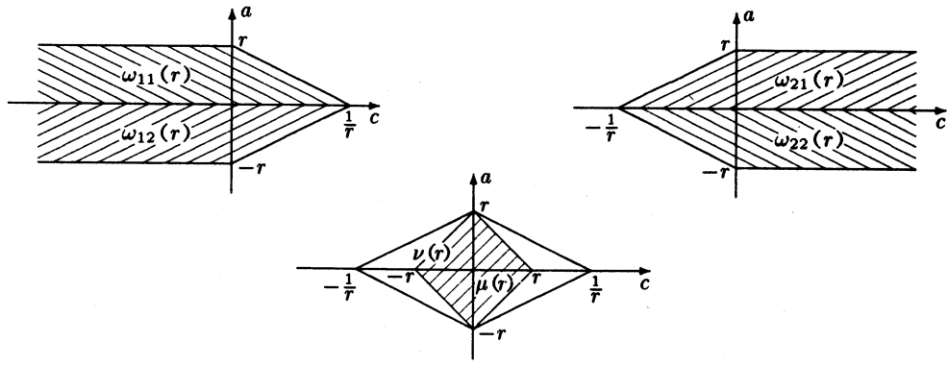

Figure 1

It is not difficult to check

Lemma 2.1. *If* $\beta_{k-1} \in [0,(-1)^i r]$ *and* $(c_k, a_k) \in \omega_{ij}(r)$, *then*

$$\beta_k \in [0,(-1)^j r], \quad i,j = 1,2, \quad k = 2,3\ldots .$$

Corollary 2.1. *If* $(c_k, a_k) \in \omega_{11}(r)$, *i.e.*,

$$0 \le a_k \le r, \quad -r^2 c_k - a_k + r \ge 0, \quad c_k \ne \frac{1}{r}, \quad k = 1,2,\ldots, \tag{2.6}$$

then $\beta_k \in [-r,0]$, $k = 1,2,\ldots .$

Corollary 2.2. *If* $(c_k, a_k) \in \omega_{22}(r)$, *i.e.*,

$$-r \le a_k \le 0, \quad r^2 c_k + a_k + r \ge 0, \quad c_k \ne -\frac{1}{r}, \quad k = 1,2,\ldots, \tag{2.7}$$

then $\beta_k \in [0,r]$, $k = 1,2,\ldots .$

Corollary 2.3. *If* $(c_k, a_k) \in \nu(r)$, *i.e.*,

$$|c_k| \ne \frac{1}{r}, \quad r^2|c_k| + |a_k| \le r, \quad k = 1,2,\ldots, \tag{2.8}$$

then $|\beta_k| \le r$, $k = 1,2,\ldots .$

Remark 2.1. For $r < 1$ we have a strict inclusion $\mu(r) \subset \nu(r)$. Thus, Corollary 2.3 is the promised weakening of condition (2.3). Moreover, it is already impossible to widen the domain $\nu(r)$ keeping $|\beta_k| \le r$. In other words, the domain $\nu(r)$ and the estimate $|\beta_k| \le r$ are the best, i.e., for any $m = 2,3,\ldots$, and for any point $(c,a) \notin \nu(r)$ the recurrent relation of the kind (2.1) exists such that

$$(c_i, a_i) \in \nu(r), \quad i < m \quad (c_m, a_m) = (c, a), \quad |\beta_m| > r.$$

Remark 2.2. The domains ω_{11}, ω_{22} in Corollaries 2.1, 2.2 and the estimates for β_k are also the best possible.

3. Boundedness of β_k. The inverse problem

Let us examine (2.1) again. Consider $a_k c_k \neq 0$ for $k > 1$. In this section we will find the sufficient conditions for the boundedness of sequence β_k and the value $\sup_k |\beta_k|$.

At first, let us find the necessary and sufficient conditions of existence of the number $r > 0$ such that all points (c_k, a_k) are in the domain $\nu(r)$. Then, if a set of such r is not empty we will find $r_0 = \inf r$. And according to Corollary 2.3, if all $(c_k, a_k) \in \nu(r_0)$, then $|\beta_k| \leq r_0$ for any k.

From (2.5) it follows that $(c_k, a_k) \in \nu(r)$ for all k if and only if

$$r^2|c_k| + |a_k| \leq r, \quad k = 1, 2, \ldots . \tag{3.1}$$

For the solvability with respect to r of the system of quadratic inequalities (3.1) it is necessary

$$|c_k a_k| \leq \frac{1}{4}, \quad k = 2, 3, \ldots . \tag{3.2}$$

Let (3.2) be satisfied. Then for any number r (3.1) has solution if and only if

$$u = \max_{k \geq 2} u_k \leq v = \min_{k \geq 2} v_k, \qquad |a_1| \leq v. \tag{3.3}$$

Here

$$u_k = \frac{1 - \sqrt{1 - 4|c_k a_k|}}{2|c_k|}, \quad v_k = \frac{1 + \sqrt{1 - 4|c_k a_k|}}{2|c_k|}, \quad k = 2, 3, \ldots .$$

For this the minimum $r = r_0$, satisfying the system (3.1) is $r_0 = \max\{|a_1|, u\}$. Thus, the following theorem takes place.

Theorem 3.1. *Let for relation* (2.1) *the following conditions be satisfied:*

1) $0 < |c_k a_k| \leq \frac{1}{4}, \quad k \geq 2;$

2) $u = \max\limits_{k \geq 2} \dfrac{1 - \sqrt{1 - 4|c_k a_k|}}{2|c_k|} \leq v = \min\limits_{k \geq 2} \dfrac{1 + \sqrt{1 - 4|c_k a_k|}}{2|c_k|};$

3) $|a_1| \leq v.$

Then $|\beta_k| \leq r_0 = \max\{|a_1|, u\}$, $k \geq 1$, *and in addition it is impossible to lower the estimate* r_0 *in the class of relations of the kind* (2.1) *satisfying conditions* 1)–3).

The similar considerations can be conducted for the domains $\omega_{11}(r)$ and $\omega_{22}(r)$. Denoting

$$I_1 = \{k | c_k a_k > 0\}, \quad I_2 = \{k | c_k a_k < 0\},$$

we arrive at the following theorem.

Theorem 3.2. *Let for relation* (2.1) $I_1, I_2 \neq \phi$ *and the following conditions be satisfied*

1) *all* a_k *have the same signs*;

2) $0 < c_k a_k \leq \frac{1}{4}, \quad k \in I_1$;

3) $u = \max\limits_{k \in I_1} \dfrac{1 - \sqrt{1 - 4c_k a_k}}{2|c_k|} \leq v = \min\limits_{k \in I_1} \dfrac{1 + \sqrt{1 - 4c_k a_k}}{2|c_k|}$;

4) $a = \max_{k \in I_2} |a_k| \leq v$.

Then $|\beta_k| \leq r_0$, *or more precisely,*

$$\beta_k \in (0, r_0 \cdot \operatorname{sign} a_k], \quad k \geq 1, \quad r_0 = \max\{|a_1|, u\}.$$

Moreover, it is impossible to lower the estimate r_0.

The question about existence of the domain $\nu(r)$, covering all setting points (c_k, a_k) has an interesting geometrical interpretation. Straight lines $\pm r^2 c \pm a = r$ from (2.5), which limit the rhomb $\nu(r)$ are considered to be tangents towards the branches of the hyperbola $|ca| = 1/4$ from (3.2).

The rhomb area $\nu(r)$ is equal to 2 and does not depend on r. The following circumstance is turned out to be somewhat unexpected. Let us take two such rhombs as $\nu(0,5)$ and $\nu(0,7)$. Neither the first nor the second rhomb lies in another. It is not difficult to imagine the set $\{(c_k, a_k) | k = 1, 2, \ldots, \}$, which is not included into $\nu(0,7)$ but lies in $\nu(0,5)$. Corollary 2.3 does not allow to confirm that all $|\beta_k| \leq 0.7$. But at the same time this very consequence guarantees that all $|\beta_k| \leq 0.5$.

Example 3.1. Let $a_1 = -0.1$, $c_k = 1000$, $a_k = -10.1$, $k = 2, 3, \ldots$. Then $\beta_k \equiv 0,1$. For this $|c_k a_k| > 1/4$ and $(c_k, a_k) \notin \nu(r)$ for any $r > 0$, $k = 2, 3, \ldots$.

This example points out that the condition $|c_k a_k| \leq 1/4$ is not necessary for $|\beta_k| \leq r$. It is necessary for hitting all (c_k, a_k) in any rhomb $\nu(r)$. And according to Corollary 2.3 the latter condition is sufficient for $|\beta_k| \leq r$.

4. Recurrent relation for $\delta\beta_k$

Let us return to non-normalized relation (1.1). We will consider $a_k c_k \neq 0$, $k \geq 2$. Investigate the process of accumulation of the round-off error in computer realization of (1.1). We will use the model

$$fl(x * y) = (x * y)(1 + \varepsilon), \quad |\varepsilon| \leq \varepsilon_M, \quad * \in \{+, -, \times, /\}. \tag{4.1}$$

Unlike the exact values β_k we will denote the values received in process of computations according to (1.1) in model (4.1) as $\tilde{\beta}_k$. Let us deduce the recurrent relation for the relative error

$$\delta\beta_k = \frac{\tilde{\beta}_k - \beta_k}{\beta_k}; \quad k = 1, 2, \ldots .$$

In accordance with (1.1), (4.1)

$$\tilde{\beta}_k = \frac{-a_k(1 + \varepsilon_{k3})}{[c_k \tilde{\beta}_{k-1}(1 + \varepsilon_{k1}) + b_k](1 + \varepsilon_{k2})},$$

$$|\varepsilon_{ki}| \leq \varepsilon_M; \quad i = 1, 2, 3; \quad k = 1, 2, \ldots .$$

Hence,

$$\tilde{\beta}_k = \frac{\beta_k\left(1 + \frac{\varepsilon_{k3} - \varepsilon_{k2}}{1 + \varepsilon_{k2}}\right)}{-q_k(1 + \varepsilon_{k1})\delta\beta_{k-1} + 1 - q_k\varepsilon_{k1}},$$

where

$$q_k = \frac{c_k}{a_k}\beta_k\beta_{k-1}, \quad k = 1; 2, \ldots . \tag{4.2}$$

Having denoted

$$\varepsilon'_M = \frac{\varepsilon_M}{1 - \varepsilon_M}, \quad q'_k = q_k(1 + \varepsilon_{k1}), \tag{4.3}$$

we obtain

$$\delta\beta_k = \frac{q'_k\delta\beta_{k-1} + q_k\varepsilon_{k1} + 2\varepsilon'_k}{-q'_k\delta\beta_{k-1} + 1 - q_k\varepsilon_{k1}}, \quad |\varepsilon'_k| \leq \varepsilon'_M, \quad k = 1, 2, \ldots . \tag{4.4}$$

5. Recurrent relations for q_k

The value q_k is the main characteristic in (4.4). From (1.1), (4.2) follows:

$$q_k = \frac{-c_k\beta_{k-1}}{c_k\beta_{k-1} + b_k}, \quad k = 2,3,\dots .$$

Now it is easy to check that

$$q_1 = 0, \quad q_k = \frac{1}{\frac{1}{d_k(1+q_{k-1})} - 1}, \tag{5.1}$$

where

$$d_k = \frac{c_k a_{k-1}}{b_k b_{k-1}}, \quad k = 2,3,\dots .$$

Iterating (5.1) we obtain

$$q_1 = 0, \quad q_2 = \frac{1}{\frac{1}{d_2} - 1}, \quad q_k = \frac{d_k}{1 - d_k - d_{k-1}q_{k-2}}, \quad k = 3,4,\dots . \tag{5.2}$$

6. Boundedness of q_k. The direct problem

Let some $q \in (0,1)$ be given. Our aim is to point out the non-recurrent conditions on the coefficients d_k sufficient for $|q_k| \le q$, $k = 2,3,\dots$.

Denote the intervals in R

$$\xi_{11}(q) = \left[\frac{-q}{1-q}, 0\right), \quad \xi_{12}(q) = \left(0, \frac{q}{1+q}\right],$$

$$\xi_{21}(q) = \left[\frac{-q}{1-q^2}, 0\right), \quad \xi_{22}(q) = \left(0, \frac{q}{(1+q)^2}\right].$$

It is not difficult to check

Lemma 6.1. *If* $q_{k-1} \in (0,(-1)^i q]$ *and* $d_k \in \xi_{ij}(q)$, *then* $q_k \in (0,(-1)^j q]$, $i,j = 1,2$, $k = 3,4,\dots$.

Corollary 6.1. *If* $d_k \in \xi_{11}(q)$, $k = 2,3,\dots$, *then* $q_k \in [-q,0)$, $k = 2,3,\dots$.

Corollary 6.2. *If* $d_2 \in \xi_{12}(q)$, $d_k \in \xi_{22}(q)$, $k = 3,4,\dots$, *then* $q_k \in (0,q]$, $k = 2,3,\dots$.

Corollary 6.3. *If* $d_2 \in \xi_{11}(q) \cup \xi_{12}(q)$, $d_2 \in \xi_{21}(q) \cup \xi_{22}(q)$, $k = 3,4,\dots$, *then* $|q_k| \le q$, $k = 2,3,\dots$.

For $q = 1$ we obtain

Lemma 6.2. *If*

$$d_2 \le 1/2, \quad d_k \le 1/4, \quad k = 3, 4, \ldots, \tag{6.1}$$

then $|q_k| \le 1$, $k = 2, 3, \ldots$.

Having applied to (5.2) the method from Section 2 we obtain, for example,

Theorem 6.1. *If*

$$d_3, d_2 \le \frac{1}{2}, \quad \frac{d_3}{1 - d_2} \le 0, \quad d_k + d_{k-1} \le \frac{1}{2}, \quad d_k \le \frac{1}{2}, \quad k = 4, 5, \ldots, \tag{6.2}$$

then $|q_k| \le 1$, $k = 2, 3, \ldots$.

In Figure 2 we show how Theorem 6.1 generalizes Lemma 6.2. Here the domain from Lemma 6.2 is covered with double hatching.

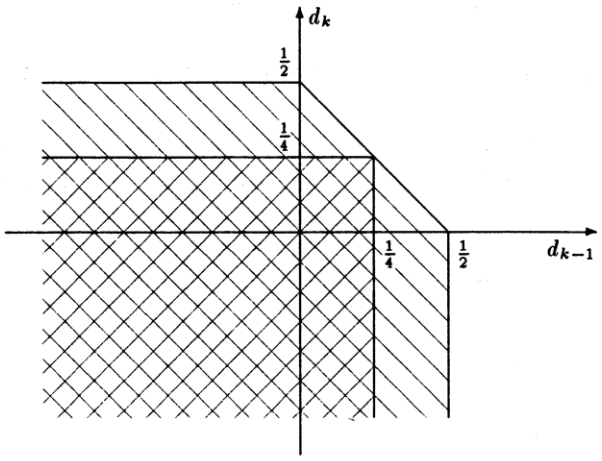

Figure 2

Remark 6.1. All domains for d_k, d_{k-1} and estimates for q_k are the best.

7. Boundedness of q_k. The inverse problem

Let there exists a concrete relation of the kind (1.1). Denote

$$d_2 = \inf d_k, \quad D = \sup d_k, \quad k \ge 3.$$

If

$$d_2 \geq 1/2, \quad D \leq 1/4, \tag{7.1}$$

then according to Lemma 6.2 $Q = \sup_{k\geq 2} |q_k| \leq 1$.

In this section we will determine more exactly the estimate for Q in conditions (7.1), having expressed it only through d and D.

Only three cases are possible.

1) $d_k < 0$, $k = 2,3,\ldots$. Then according to Corollary 6.1 all $q_k < 0$ and all d_k should be considered in $\xi_{11}(q)$. Now it is necessary to define the minimum $Q_1 > 0$ such that d_2, $d \in \xi_{11}(Q_1)$. It is evident that this is

$$Q_1 = \max\left\{\frac{-d_2}{1-d_2}, \frac{-d}{1-d}\right\} < 1.$$

2) $d_k > 0$, $k = 2,3,\ldots$. According to Corollary 6.2 all $q_k > 0$ and we ought to find minimum $Q_2 > 0$ such that $d_2 \in \xi_{12}(Q_2)$, $D \in \xi_{22}(Q_2)$. It is evident that this is

$$Q_2 = \max\left\{\frac{|d_2|}{1-|d_2|}, \frac{1-2D-\sqrt{1-4D}}{2D}\right\} \leq 1.$$

3) There exists d_k of different signs. According to Corollary 6.3 we ought to find the minimum $Q_3 > 0$ such that $d_2 \in \xi_{11}(Q_3) \cup \xi_{12}(Q_3)$, $d, D \in \xi_{21}(Q_3) \cup \xi_{22}(Q_3)$. This is

$$Q_3 = \max\left\{\frac{|d_2|}{1-d_2}, \frac{1-\sqrt{1+4d^2}}{2d}, \frac{1-2D-\sqrt{1-4D}}{2D}\right\} \leq 1.$$

Remark 7.1. In the class of relations of the kind (1.1) satisfying (7.1) the estimates Q_1, Q_2, Q_3 are attained.

Now one more estimate for $|\beta_k|$ can be done. From (1.1), (4.2) it follows:

$$|\beta_k| = \left|\frac{a_k}{b_k}(1+q_k)\right| \leq \left|\frac{a_k}{b_k}\right|(1+Q_i), \quad k \geq 2,$$

for all these $i = 1,2,3$, which are chosen according to which case out of the three ones takes place.

At last, in condition $d_2 \leq 1/2$, $d_k \leq 1/4$, $k \geq 3$ we will get the lower estimate for $|c_k\beta_{k-1} + b_k|$. From the latter estimate for β_k and Lemma 6.2 it follows:

$$|c_k\beta_{k-1} + b_k| = |1 - d_k(1+q_{k-1})||b_k| \geq 0,5|b_k|, \quad k \geq 2.$$

In other words, if $|b_2| > \varepsilon$, $|b_k| > 2\varepsilon$, $k \geq 3$, then the absolute value of all denominators in (1.1) are more than ε.

8. The estimates of $\delta\beta_k$

Now we can finish the investigation started in Section 4. In Sections 6, 7 the estimates for $Q = \sup|q_k|$ are given. Let $Q \leq 1$. Then from (4.4) follows:

$$|\delta\beta_k| \leq \frac{Q'|\delta\beta_{k-1}| + 3\varepsilon'_M}{-Q'|\delta\beta_{k-1}| + 1 - Q\varepsilon_M}, \quad Q' = Q(1+\varepsilon_M), \tag{8.1}$$

for those numbers $k = 2, 3, \ldots, s$, for which the denominator in (8.1) will remain positive. The estimate s will be obtained below. Consider the right-hand side of (8.1) as a fraction-linear function $\mathcal{L}(|\delta\beta_{k-1}|)$ with the matrix

$$L = \begin{bmatrix} Q' & 3\varepsilon'_M \\ -Q' & 1 - Q\varepsilon_M \end{bmatrix}.$$

This function is the increasing one. Therefore,

$$|\delta\beta_{k+1}| \leq \mathcal{L}(|\delta\beta_k|) \leq \mathcal{L}(\mathcal{L}(|\delta\beta_{k-1}|)) \leq \ldots \leq \mathcal{L}^k(|\delta\beta_1|) \leq \mathcal{L}^k(\varepsilon_M).$$

The function l^k is again fraction-linear with the matrix

$$l^k = \begin{bmatrix} l_{11}^{(k)} & l_{12}^{(k)} \\ l_{21}^{(k)} & l_{22}^{(k)} \end{bmatrix}.$$

Then

$$|\delta\beta_{k+1}| \leq \mathcal{L}^k(\varepsilon_M) = \frac{l_{11}^{(k)}\varepsilon_M + l_{12}^{(k)}}{l_{21}^{(k)}\varepsilon_M + l_{22}^{(k)}}, \tag{8.1}$$

for those numbers k, for which the denominator here remains positive.

Having raised the matrix L to power k by induction we obtain

$$\begin{aligned} l_{11}^{(k)}\varepsilon_M + l_{12}^{(k)} &\leq Q'^k\varepsilon_M + 3(Q'^{k-1} + \cdots + Q' + 1)\varepsilon'_M, \\ l_{21}^{(k)}\varepsilon_M + l_{22}^{(k)} &\geq 1 - (Q'^k + \cdots + Q'^2 + Q' + kQ)\varepsilon_M - \\ &\quad 3(Q'^k + 2Q'^{k-1} + \cdots + (k-1)Q')\varepsilon_M. \end{aligned}$$

Hence, from (8.2) for $Q' \leq 1$ follows:

$$|\delta\beta_{k+1}| \leq \frac{(3k+1)\varepsilon'_M}{1 - \left(\frac{3}{2}k^2 + \frac{1}{2}k\right)\varepsilon'_M}, \quad k \leq s, \tag{8.2}$$

where s is the maximum, for which the denominator is positive here, i.e.,

$$s \le \sqrt{\frac{2}{3\varepsilon'_M}} - 1.$$

Similarly, for $Q' < 1$

$$|\delta\beta_{k+1}| \le \frac{4\varepsilon'_M}{1 - Q' - 4kQ'\varepsilon'_M}, \quad k < \frac{1 - Q'}{4Q'\varepsilon'_M}. \tag{8.3}$$

Note that the inequalities for $|\delta\beta_{k+1}|$ in (8.3), (8.4) are valid, but not "with accuracy, till small higher orders". These inequalities do not contain any indefinite constants and do not contain, in the evident form, the conditionalities of the matrix tridiag (c_k, b_k, a_k). The constant Q' from estimates (8.3), (8.4), *a priori* is estimated easily by the methods of Sections 6, 7.

Example 8.1. Let

$$a_1 = \frac{1 - \sqrt{241}}{20}, \quad a_k = -6, \quad b_1 = b_k = 1, \quad c_k = 10, \quad k = 2, 3, \ldots .$$

Then

$$\beta_k \equiv \beta_1 = -a_1 \in (0, 1), \quad Q = Q_1 = 60/61 < 1.$$

It means that (8.4) takes place.

References

[1] V.P. Il'in, Yu.I. Kuznetsov, Three-diagonal Matrices and their Applications, Nauka, Moskow, 1985 (in Russian).