

## *Remarks to Accelerated Iterative Processes for Numerical Solution of Equations*

Masatomo FUJII

(Received September 1, 1963)

### 1. Introduction.

In the present paper, we are concerned with error estimation for some accelerated iterative processes for numerical solution of equations, the round-off errors committed in actual computation being taken into consideration. From our results, some remarks will be made on the use of these formulas in actual computation.

For simplicity, we suppose the equation is given in the form

$$(1.1) \quad x = \varphi(x),$$

where  $\varphi(x)$  is continuously differentiable in the closed interval  $I$ .

We assume that

$$(1.2) \quad 0 \leq |\varphi'(x)| \leq K < 1 \quad \text{in } I$$

and that

$$(1.3) \quad S \{h: |h - x_1| \leq \frac{K}{1-K} |x_1 - x_0|\} \subset I$$

for  $x_0 \in I$  and  $x_1 = \varphi(x_0)$ . Then it is well known [2] that

1° the iterative process

$$(1.4) \quad x_{n+1} = \varphi(x_n) \quad (n = 0, 1, 2, \dots)$$

can be continued indefinitely so that  $x_n \in S$  ( $n=1, 2, \dots$ );

2° the sequence  $\{x_n\}$  ( $n=0, 1, 2, \dots$ ) converges in  $S$  and  $\lim_{n \rightarrow \infty} x_n = \bar{x}$  satisfies the equation (1.1);

3°  $x = \bar{x}$  is the unique solution of (1.1) in  $I$ .

As is well known, the convergence of this classical iterative process (1.4) (in what follows this process is abbreviated as CI-process) is not fast when  $K$  is not small. To rescue this fault, some methods are devised to accelerate the speed of the convergence of the above iterative process.

One of these accelerated processes is Aitken's  $\delta^2$  process [1, 3, 4, 5]. It is based on using the function

$$(1.5) \quad \psi(x) = \varphi\{\varphi(x)\} - \frac{[\varphi\{\varphi(x)\} - \varphi(x)]^2}{\varphi\{\varphi(x)\} - 2\varphi(x) + x}$$

in place of the given function  $\varphi(x)$ . One process is to predict the root  $\bar{x}$  by

$$(1.6) \quad \bar{x}_{n+2} = \psi(x_n),$$

where

$$(1.7) \quad x_{i+1} = \varphi(x_i) \quad (i = 0, 1, 2, \dots, n-1).$$

In what follows, we shall call this process Aitken's predictive process or briefly the AP-process. Another process is the iterative process:

$$(1.8) \quad \tilde{x}_{n+1} = \psi(\tilde{x}_n) \quad (n = 0, 1, 2, \dots).$$

This process will be called Aitken's iterative process or briefly the AI-process in what follows.

Aitken's processes can be modified by using the function

$$(1.9) \quad \mathcal{P}(x) = x - \frac{\varphi(x) - x}{\varphi'(x) - 1}$$

in place of  $\psi(x)$ . This modification is suggested from the fact that

$$\begin{aligned} \psi(x) &= x - \frac{\{\varphi(x) - x\}}{\varphi\{\varphi(x)\} - 2\varphi(x) + x} \{\varphi(x) - x\} \\ &\approx x - \frac{1}{\varphi'(x) - 1} \{\varphi(x) - x\} \end{aligned}$$

for  $x \approx \bar{x}$ . Corresponding to the processes (1.6) and (1.8), there are conceived the predictive process

$$(1.10) \quad \bar{x}_{n+1} = \mathcal{P}(x_n)$$

where

$$(1.11) \quad x_{i+1} = \varphi(x_i) \quad (i = 0, 1, 2, \dots, n-1)$$

and the iterative process

$$(1.12) \quad \tilde{x}_{n+1} = \mathcal{P}(\tilde{x}_n) \quad (n = 0, 1, 2, \dots).$$

In what follows, these are called the modified Aitken's predictive process (MAP-process for short) and the modified Aitken's iterative process (MAI-process for short) respectively. Evidently the MAI-process is nothing but Newton's process applied to the equation

$$(1.13) \quad \varphi(x) - x = 0.$$

The above modified Aitken's processes can be simplified in analogous way

as in Newton's process. Namely using the function

$$(1.14) \quad \Phi(x) = x - k \{\varphi(x) - x\}$$

in place of  $\Psi(x)$ , where  $k$  is an arbitrary number such that

$$(1.15) \quad k \approx \frac{1}{\varphi'(\tilde{x}) - 1},$$

there are conceived the predictive process

$$(1.16) \quad \bar{x}_{n+1} = \Phi(x_n)$$

where

$$(1.17) \quad x_{i+1} = \varphi(x_i) \quad (i = 0, 1, 2, \dots, n-1)$$

and the iterative process

$$(1.18) \quad \tilde{x}_{n+1} = \Phi(\tilde{x}_n) \quad (n = 0, 1, 2, \dots).$$

In what follows, the former is called the simplified Aitken's predictive process (SAP-process for short) and the latter the simplified Aitken's iterative process (SAI-process for short). The latter process is nothing but the simplified Newton's process applied to the equation (1.13).

The geometrical meanings of the above accelerated processes can be readily seen from Figs. 1-3.

In the present paper, assuming

$$(1.19) \quad |\varphi'(x_1) - \varphi'(x_2)| \leq L |x_1 - x_2|$$

for any  $x_1, x_2 \in I$ , we shall seek the error estimates for the approximate roots

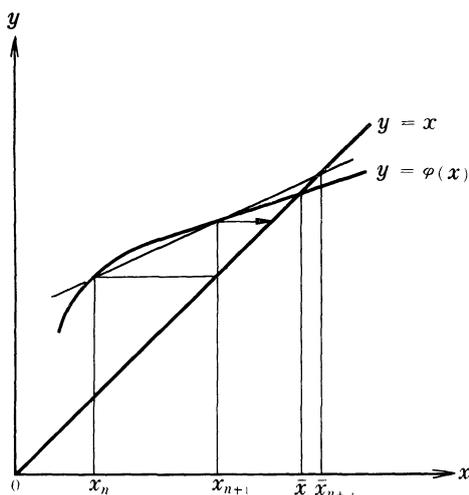


Fig. 1 AP-process

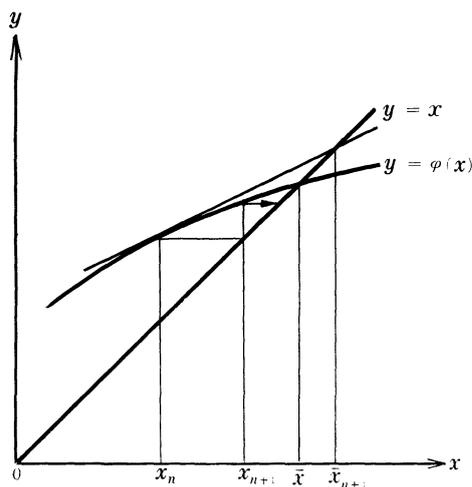


Fig. 2 MAP-process

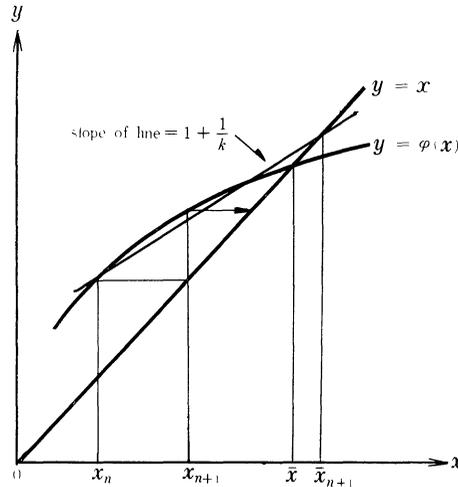


Fig. 3 SAP-process

of the given equation (1.1) obtained actually by the above accelerated processes. And, from these error estimates, some remarks will be made about using the accelerated processes for numerical computation of the roots of the equation.

In the next paragraph, in preparation for subsequent discussions, there will be described briefly the results of Urabe [6, 7] concerning the actual CI-process, namely the CI-process accompanied with the errors unavoidable in actual computation of  $\varphi(x)$ .

In the subsequent paragraph, the accelerated processes described above will be discussed in order.

## 2. The results of Urabe on the actual CI-process.

Let  $\varepsilon$  be the bound of errors committed in computation of  $\varphi(x)$ . Due to the errors in computation of  $\varphi(x)$ , there is obtained the sequence  $\{x_n^*\}$  in the actual CI-process and this sequence  $\{x_n^*\}$  differs in general from the sequence  $\{x_n\}$  obtained by the ideal CI-process (1.4).

Let us write

$$(2.1) \quad x_{n+1}^* = \varphi^*(x_n^*) = \varphi(x_n^*) + \varepsilon_n \quad (n = 0, 1, 2, \dots),$$

then, by our assumption, it is evident that

$$(2.2) \quad |\varepsilon_n| \leq \varepsilon.$$

The results of Urabe are as follows.

Assume

$$(2.3) \quad \sum \{h: |h - x_1^*| \leq \frac{K}{1-K} |x_1^* - x_0^*| + 2\delta\} \subset I,$$





$$(3.3) \quad k_n = \frac{x_{n+2}^* - x_{n+1}^*}{x_{n+2}^* - 2x_{n+1}^* + x_n^*},$$

and  $\eta_n$  is a round-off error caused by computation of  $k_n(x_{n+2}^* - x_{n+1}^*)$ .

Let us suppose the initial CI-process (2.1) is stopped by the criterion of the form (2.7) and there holds an inequality

$$(3.4) \quad 0 < \beta \leq |x_{n+1}^* - x_n^*| \leq \alpha$$

actually for the computed values  $x_n^*$  and  $x_{n+1}^*$ . Then, by §2, the inequalities (2.6) hold for the present  $x_n^*$  and  $x_{n+1}^*$ .

Now, from (3.1) and (3.2), it holds that

$$(3.5) \quad \begin{aligned} \bar{x}_{n+2}^* - \bar{x} &= \{(1 - k_n)\varphi'(\xi) + k_n\} (x_{n+1}^* - \bar{x}) \\ &\quad + (1 - k_n)\varepsilon_{n+1} + \eta_n, \end{aligned}$$

where  $\xi \in (\bar{x}, x_{n+1}^*)$ . Here the symbol  $(\bar{x}, x_{n+1}^*)$  means the open interval with the end points  $\bar{x}$  and  $x_{n+1}^*$  regardless of their magnitude.

Let us rewrite  $k_n$  as follows:

$$(3.6) \quad k_n = \frac{\lambda_n}{\lambda_n - 1},$$

where

$$(3.7) \quad \lambda_n = \frac{x_{n+2}^* - x_{n+1}^*}{x_{n+1}^* - x_n^*} = \frac{\varepsilon_{n+1} - \varepsilon_n}{x_{n+1}^* - x_n^*} + \varphi'(\xi_1)$$

and  $\xi_1 \in (x_n^*, x_{n+1}^*)$ . Then, from (1.2) and (3.4), it follows readily that

$$0 \leq 1 - k_n = \frac{1}{1 - \lambda_n} \leq \frac{1}{1 - K - \frac{2\varepsilon}{\beta}},$$

provided  $\varepsilon \ll \beta$ .

On the other hand, by (1.19), (2.6) and (3.4), it holds that

$$|\varphi'(\xi) - \varphi'(\xi_1)| \leq L \frac{\alpha + \varepsilon}{1 - K}.$$

Then it follows that

$$\begin{aligned} & |(1 - k_n)\varphi'(\xi) + k_n| \\ &= (1 - k_n) \left| \varphi'(\xi) - \varphi'(\xi_1) - \frac{\varepsilon_{n+1} - \varepsilon_n}{x_{n+1}^* - x_n^*} \right| \\ &\leq \frac{1}{1 - K - \frac{2\varepsilon}{\beta}} \left[ \frac{L}{1 - K} \alpha + \frac{L}{1 - K} \varepsilon + \frac{2\varepsilon}{\beta} \right]. \end{aligned}$$

Thence, using (2.6), from (3.5), we obtain

$$(3.8) \quad |\bar{x}_{n+2}^* - \bar{x}| \leq \frac{1}{1-K-\frac{2\varepsilon}{\beta}} \left[ \frac{L}{1-K} \alpha + \frac{L}{1-K} \varepsilon + \frac{2\varepsilon}{\beta} \right] \\ \times \left( \frac{K\alpha}{1-K} + \frac{\varepsilon}{1-K} \right) + \frac{1}{1-K-\frac{2\varepsilon}{\beta}} \varepsilon + |\eta_n|.$$

The assumption that  $\varepsilon \ll \beta$  is a natural one, for,

1° otherwise, in computation  $k_n$ , many significant figures may be lost for the denominator and there may arise an overflow in the computer as will be shown in the example in the end of the present paragraph;

2° the AP-process is expected to be more effective for prediction of the solution than the initial CI-process and this means in the actual computation that  $\bar{x}_{n+2}^*$  is exact enough even if  $x_{n+2}^*$  is not so exact or, in other words,  $|x_{n+1}^* - x_n^*|$  is not so small.

From such a point of view, in what follows, we assume

$$(3.9) \quad 1 \gg \alpha \approx \beta \gg \varepsilon.$$

Then (3.8) can be written approximately as follows:

$$(3.10) \quad |\bar{x}_{n+2}^* - \bar{x}| \leq \frac{LK}{(1-K)^3} \alpha^2 + \frac{1+K}{(1-K)^2} \varepsilon + |\eta_n|.$$

This is the desired error estimate for  $\bar{x}_{n+2}^*$  obtained actually by the AP-process.

Now, from (3.6) and (3.7), it is evident that

$$|k_n| \leq \frac{K + \frac{2\varepsilon}{\beta}}{1-K-\frac{2\varepsilon}{\beta}} \approx \frac{K}{1-K}.$$

Further, from (2.1) and (3.4), it is evident that

$$|x_{n+2}^* - x_{n+1}^*| \leq 2\varepsilon + K\alpha.$$

Therefore  $\eta_n$  is of the same magnitude as  $\varepsilon$ .

Then, from (3.10), it is seen that

$$(3.11) \quad \bar{x}_{n+2}^* - \bar{x} = O(\varepsilon)$$

if  $\alpha$  is chosen so that  $\alpha = O(\varepsilon^{\frac{1}{2}})$ . Since our computation is carried out within the error bound  $O(\varepsilon)$ , we can not expect to be able to obtain the values more exact in order than  $\bar{x}_{n+2}^*$  obtained just above.

For  $x_{n+2}^*$  obtained by the CI-process (2.1), it is readily seen from (2.6) that

$$(3.12) \quad |x_{n+2}^* - \bar{x}| \leq \frac{K^2}{1-K} \alpha + \frac{1}{1-K} \varepsilon.$$

The right member of this inequality is  $O(\alpha)$  due to (3.9). Consequently it is evident that

$$(3.13) \quad |x_{n+2}^* - \bar{x}| \gg |\bar{x}_{n+2}^* - \bar{x}|,$$

which implies the AP-process really accelerates the iterative process in the actual computation.

From (3.10), it is evident that

$$|\bar{x}_{n+2}^* - \bar{x}| \leq \frac{LK}{(1-K)^3} \alpha^2$$

approximately when  $\alpha \gg O(\varepsilon^{\frac{1}{2}})$ . However, even in this case, it is needless to say that (3.13) holds.

**Example.**

$$\varphi(x) = x - 0.5x^2 + 0.04.$$

We assume that the computation is always carried out correctly to 8 decimal places, and we consider the CI-process in the interval  $I[0.28, 0.30]$  starting from  $x_0 = 0.29$ . Then evidently

$$K = 0.72 \quad \text{and} \quad L = 1.$$

Since

$$\varepsilon = \frac{1}{2} (1 + 0.5) \times 10^{-8},$$

we see that

$$\delta = \frac{75}{28} \times 10^{-8} \approx 2.7 \times 10^{-8}.$$

Therefore, since  $x_1^* = x_1 = 0.28795$ , (2.3) is Valid, because

$$\frac{0.72}{0.28} \times 0.00205 + 5.4 \times 10^{-8} < 0.00795 < 0.01205.$$

By the actual computation, we have Table 1.

For  $n=9, 10, \dots$ , it is readily seen that

$$|\eta_n| \leq \frac{1}{2} \times 10^{-8} + \frac{1}{2} \times 10^{-8} \times 10^{-4} \approx \frac{1}{2} \times 10^{-8}.$$

Therefore, by (3.10), we have the error estimates  $\bar{e}_n$  for  $\bar{x}_n^*$  as is shown in Table 2. In Table 2, for comparison, the true error  $e_n$  for  $\bar{x}_n^*$  are also shown.

Table 1.

$n$	$x_n^*$	$x_{n+1}^* - x_n^*$	$\bar{x}_n^*$
9	0.2831 9592	-0.0000 9997	
10	0.2830 9595	-0.0000 7166	
11	0.2830 2429	-0.0000 5138	0.2828 4290
12	0.2829 7291	-0.0000 3684	0.2828 4274
13	0.2829 3607	-0.0000 2641	0.2828 4273
14	0.2829 0966	-0.0000 1894	0.2828 4279
15	0.2828 9072	-0.0000 1358	0.2828 4270
16	0.2828 7714	-0.0000 0974	0.2828 4273
17	0.2828 6740	-0.0000 0699	0.2828 4269
18	0.2828 6041	-0.0000 0501	0.2828 4264
19	0.2828 5540	-0.0000 0359	0.2828 4272
20	0.2828 5181	-0.0000 0258	0.2828 4273
21	0.2828 4923	-0.0000 0185	0.2828 4264
	⋮	⋮	⋮
	⋮	⋮	⋮
	⋮	⋮	⋮
38	0.2828 4273	-0.0000 0001	overflow
39	0.2828 4272	0.0000 0000	overflow
40	0.2828 4272	0.0000 0000	0.2828 4272

True value  $\bar{x}=0.2828\ 42712\dots$

Table 2.

$n$	$10^8 \times \bar{e}_n$	$10^8 \times e_n$
11	50	19
12	34	3
13	26	2
14	21	8
15	19	-1
16	18	2
17	18	-2
18	17	-7
19	17	1
20	17	2
21	17	-7

From Table 2, we see that the error estimates given by (3.10) are considerably good and that the values  $\bar{x}_{11}^*, \bar{x}_{12}^*, \dots, \bar{x}_{21}^*$  are accurate enough though the number of steps in the CI-process is not large and the values  $x_{11}^*, x_{12}^*, \dots, x_{21}^*$  obtained by the CI-process are far from the true value.

Table 3 shows that the overflow arises actually in the computer for computation of  $\bar{x}_{36}^*, \bar{x}_{38}^*, \bar{x}_{39}^*$  and the indefinite form appears for  $\bar{x}_{41}^*, \bar{x}_{42}^*, \dots$ .

Table 3.

$n$	$x_n^*$	$x_{n+1}^* - x_n^*$	$\bar{x}_n^*$
33	0.2828 4282	-0.0000 0003	0.2828 4269
34	0.2828 4279	-0.0000 0002	0.2828 4274
35	0.2828 4277	-0.0000 0002	0.2828 4273
36	0.2828 4275	-0.0000 0001	overflow
37	0.2828 4274	-0.0000 0001	0.2828 4273
38	0.2828 4273	-0.0000 0001	overflow
39	0.2828 4272	0.0000 0000	overflow
40	0.2828 4272	0.0000 0000	0.2828 4272
41	0.2828 4272	0.0000 0000	indefinite
	⋮	⋮	⋮
	⋮	⋮	⋮
	⋮	⋮	⋮

**Remark.** As is seen from the results of the present paragraph, in actu-

al application of the AP-process, it should be kept in mind that we should not repeat the initial CI-process too many times. *It is most desirable for effective use of the AP-process to stop the initial CI-process in the state that  $|x_{n+1}^* - x_n^*| = O(\varepsilon^{\frac{1}{2}})$ .*

#### 4. The actual AI-process.

By (1.8), the AI-process is carried out in the actual computation as follows:

$$(4.1) \quad \tilde{x}_{n+1}^* = \varphi^* \{ \varphi^*(\tilde{x}_n^*) \} - \tilde{k}_n [ \varphi^* \{ \varphi^*(\tilde{x}_n^*) \} - \varphi^*(\tilde{x}_n^*) ] + \eta_n,$$

where

$$(4.2) \quad \tilde{k}_n = \frac{\varphi^* \{ \varphi^*(\tilde{x}_n^*) \} - \varphi^*(\tilde{x}_n^*)}{\varphi^* \{ \varphi^*(\tilde{x}_n^*) \} - 2\varphi^*(\tilde{x}_n^*) + \tilde{x}_n^*}$$

and  $\eta_n$  is a round-off error caused by computation of  $\tilde{k}_n [ \varphi^* \{ \varphi^*(\tilde{x}_n^*) \} - \varphi^*(\tilde{x}_n^*) ]$ .

Since  $\tilde{k}_n$  is of the same form as  $k_n$ , as is remarked in the preceding paragraph, the iterative process should be stopped before  $| \varphi^*(\tilde{x}_n^*) - \tilde{x}_n^* |$  becomes so small that it may be  $O(\varepsilon)$ . Hence we suppose that the AI-process under consideration is stopped in the state where

$$(4.3) \quad 0 < \beta \leq | \varphi^*(\tilde{x}_n^*) - \tilde{x}_n^* | \leq \alpha$$

and that

$$(4.4) \quad \varepsilon \ll \beta.$$

Now, as is seen from the proof of (2.6), it is valid that

$$(4.5) \quad \begin{cases} |x^* - \bar{x}| \leq \frac{\alpha + \varepsilon}{1 - K} \\ |\varphi^*(x^*) - \bar{x}| \leq \frac{K\alpha + \varepsilon}{1 - K} \end{cases}$$

for any  $x^*$  such that

$$(4.6) \quad | \varphi^*(x^*) - x^* | \leq \alpha.$$

Then the results in the preceding paragraph are valid if  $x_n^*, x_{n+1}^*, x_{n+2}^*$  and  $\bar{x}_{n+2}^*$  are replaced by  $\tilde{x}_n^*, \varphi^*(\tilde{x}_n^*), \varphi^* \{ \varphi^*(\tilde{x}_n^*) \}$  and  $\tilde{x}_{n+1}^*$  respectively.

Thus, from (3.8) and (3.10), for the AI-process under consideration, we have

$$(4.7) \quad | \tilde{x}_{n+1}^* - \bar{x} | \leq \frac{1}{1 - K - \frac{2\varepsilon}{\beta}} \left[ \frac{L}{1 - K} \alpha + \frac{L}{1 - K} \varepsilon + \frac{2\varepsilon}{\beta} \right]$$

$$\times \left( \frac{K}{1-K} \alpha + \frac{1}{1-K} \varepsilon \right) + \frac{1}{1-K - \frac{2\varepsilon}{\beta}} \varepsilon + |\eta_n|$$

and, in case  $\alpha \approx \beta$ ,

$$(4.8) \quad |\tilde{x}_{n+1}^* - \bar{x}| \leq \frac{LK}{(1-K)^3} \alpha^2 + \frac{1+K}{(1-K)^2} \varepsilon + |\eta_n|.$$

These are the desired error estimates for the values  $\tilde{x}_{n+1}^*$  obtained by the AI-process.

As is seen from comparison of (3.10) with (3.12), the AP-process once applied yields an acceleration of the CI-process, *consequently the AI-process which is a repetition of the AP-processes yields an acceleration of the AP-process and accelerates the CI-process more than the AP-process.*

**Example.** The AI-process applied for the example of the preceding paragraph.

Table 4 shows the results of the actual computation.

Table 4.

$n$	$\tilde{x}_n^*$	$\varphi^*(\tilde{x}_n^*) - \tilde{x}_n^*$	$10^8 \times \tilde{e}_n$	$10^8 \times e_n$
0	0.2900 0000	-0.0020 5		
1	0.2829 0598	-0.0000 179	13801	6327
2	0.2828 4266	0.0000 0001	18	- 5
3	overflow			

Here  $\tilde{e}_n$  refer to the error estimates for  $\tilde{x}_n^*$  by means of (4.8) and  $e_n$  refer to the true errors for  $\tilde{x}_n^*$ .

Comparing Table 4 with Table 1, we see that the AI-process yields really an acceleration of the AP-process and accelerates the CI-process much more.

Table 4 shows also that the process should be stopped before  $|\varphi^*(\tilde{x}_n^*) - \tilde{x}_n^*|$  becomes too small.

## 5. The actual MAP- and MAI-processes.

By (1.10), the MAP-process is carried out in the actual computation as follows:

$$(5.1) \quad \bar{x}_{n+1}^* = x_n^* - \frac{x_{n+1}^* - x_n^*}{\varphi'^*(x_n^*) - 1} + \eta_n,$$

where

$$x_{i+1}^* = \varphi^*(x_i^*) \quad (i = 0, 1, 2, \dots, n),$$

and  $\varphi'^*(x_n^*)$  is a computed value of  $\varphi'(x_n^*)$  and  $\eta_n$  is a round-off error caused by computation of

$$\frac{x_{n+1}^* - x_n^*}{\varphi'^*(x_n^*) - 1}.$$

Let us suppose

$$(5.2) \quad |\varphi'^*(x_n^*) - \varphi'(x_n^*)| \leq \varepsilon' \quad (n = 0, 1, 2, \dots).$$

From (5.1), it follows readily that

$$(5.3) \quad \bar{x}_{n+1}^* - \bar{x} = (x_n^* - \bar{x}) \frac{\varphi'(\xi) - \varphi'^*(x_n^*)}{1 - \varphi'^*(x_n^*)} + \frac{\varepsilon_n}{1 - \varphi'^*(x_n^*)} + \eta_n,$$

where  $\xi \in (\bar{x}, x_n^*)$ .

Let us suppose that (2.7) is valid in the present case. Then the former of (2.6) is valid, consequently, from (5.3), readily follows:

$$(5.4) \quad |\bar{x}_{n+1}^* - \bar{x}| \leq \frac{1}{1 - K - \varepsilon'} \left[ \frac{L(\alpha + \varepsilon)}{1 - K} + \varepsilon' \right] \frac{\alpha + \varepsilon}{1 - K} + \frac{\varepsilon}{1 - K - \varepsilon'} + |\eta_n|.$$

Since  $\alpha \ll 1$  and  $\varepsilon' = O(\varepsilon)$ , neglecting the quantities smaller than  $O(\varepsilon)$ , we can write (5.4) approximately as follows:

$$(5.5) \quad |\bar{x}_{n+1}^* - \bar{x}| \leq \frac{L}{(1 - K)^3} \alpha^2 + \frac{\varepsilon}{1 - K} + |\eta_n|.$$

*This is the desired error estimate for  $\bar{x}_{n+1}^*$  obtained by the MAP-process.*

**Example 1.** The MAP-process applied for the example of §3. Here,

$$|\eta_n| \leq \frac{1}{2} \times 10^{-8}.$$

Table 5 shows the results of the actual computation.

Table 5.

$n$	$x_n^*$	$x_{n+1}^* - x_n^*$	$10^8 \times \bar{e}_n$	$10^8 \times e_n$
10		-0.0000 7166		
11	0.2828 4282	-0.0000 5138	27	11
12	0.2828 4275	-0.0000 3684	15	4
13	0.2828 4272	-0.0000 2641	9	1
14	0.2828 4273	-0.0000 1894	6	2
15	0.2828 4271	-0.0000 1358	5	0
16	0.2828 4272	-0.0000 0974	4	1
17	0.2828 4271	-0.0000 0699	4	0
18	0.2828 4269	-0.0000 0501	3	-2
19	0.2828 4270	-0.0000 0359	3	-1
20	0.2828 4271	-0.0000 0258	3	0
21	0.2828 4269		3	-2

Here  $\bar{e}_n$  refer to the error estimates given by (5.5) and  $e_n$  refer to the true errors. Table 5 shows that the error estimates given by (5.5) are considerably good and that the values  $\bar{x}_{11}^*$ ,  $\bar{x}_{12}^*$ ,  $\bar{x}_{13}^*$ ,  $\dots$ ,  $\bar{x}_{21}^*$  are accurate in the same degree as the values obtained by the AP-process (cf. Table 2).

The MAI-process is, as is remarked in §1, a Newton's process applied for the equation (1.13). Therefore the error estimates are obtained by applying the theory of Urabe [7] to the function  $\Psi(x)$  defined by (1.9).

From (1.9), it is readily seen that

$$\Psi(x) - \Psi(\bar{x}) = (x - \bar{x}) \frac{\varphi'(\xi) - \varphi'(x)}{1 - \varphi'(x)},$$

where  $\xi \in (\bar{x}, x)$ . Consequently, by the assumption (1.19), it holds that

$$(5.6) \quad |\Psi(x) - \Psi(\bar{x})| \leq K(x, \bar{x}) |x - \bar{x}|.$$

where

$$(5.7) \quad K(x, \bar{x}) = \frac{L}{1-K} |x - \bar{x}|.$$

Now let  $\Psi^*(\hat{x}_n^*)$  be the values of  $\Psi(\hat{x}_n^*)$  obtained in the actual computation. Then evidently

$$(5.8) \quad \Psi^*(\hat{x}_n^*) = \hat{x}_n^* - \frac{\varphi^*(\hat{x}_n^*) - \hat{x}_n^*}{\varphi'^*(\hat{x}_n^*) - 1} + \eta_n,$$

where  $\eta_n$  is a round-off error caused by computation of

$$\frac{\varphi^*(\hat{x}_n^*) - \hat{x}_n^*}{\varphi'^*(\hat{x}_n^*) - 1}.$$

By the way, by our assumptions (2.2) and (5.2),

$$(5.9) \quad \left| \frac{\varphi^*(\hat{x}_n^*) - \hat{x}_n^*}{\varphi'^*(\hat{x}_n^*) - 1} - \frac{\varphi(\hat{x}_n^*) - \hat{x}_n^*}{\varphi'(\hat{x}_n^*) - 1} \right| \leq \frac{1}{1-K-\varepsilon'} \varepsilon + \frac{|\varphi(\hat{x}_n^*) - \hat{x}_n^*|}{(1-K)(1-K-\varepsilon')} \varepsilon'.$$

Let us put

$$M_1 = \max_{\hat{x}_n^* \in J_1} |\varphi(\hat{x}_n^*) - \hat{x}_n^*|,$$

where

$$J_1 = \{\hat{x}_n^* : \hat{x}_n^* \text{ in the state of ONC}\},$$

and assume that

$$|\eta_n| \leq \gamma.$$

Then, from (5.8) and (5.9), it follows that, for  $\hat{x}_n^* \in J_1$ ,

$$(5.10) \quad |\Psi^*(\tilde{x}_n^*) - \Psi(\tilde{x}_n^*)| \leq \varepsilon'',$$

where

$$(5.11) \quad \varepsilon'' = \frac{1}{1-K-\varepsilon'} \varepsilon + \frac{M_1}{(1-K)(1-K-\varepsilon')} \varepsilon' + \eta.$$

Hence, by the theory of Urabe [7], it is seen that

$$(5.12) \quad |\tilde{x}_n^* - \bar{x}| \leq \frac{1-K}{2L} \left[ 1 - \sqrt{1 - \frac{4L\varepsilon''}{1-K}} \right] \approx \varepsilon''$$

for  $\tilde{x}_n^*$  in the state of ONC.

However,

$$|\varphi(x) - x| \leq |\varphi(x) - \varphi(\bar{x})| + |\bar{x} - x| \leq (1+K)|x - \bar{x}|,$$

consequently, for  $\tilde{x}_n^*$  in the state of ONC,

$$(5.13) \quad |\varphi(\tilde{x}_n^*) - \tilde{x}_n^*| \leq \frac{1-K^2}{2L} \left[ 1 - \sqrt{1 - \frac{4L\varepsilon''}{1-K}} \right] \approx (1+K)\varepsilon''.$$

Therefore, in this case, we obtain the relation

$$(5.14) \quad \varepsilon'' \leq \frac{1}{1-K-\varepsilon'} \varepsilon + \frac{(1+K)\varepsilon''}{(1-K)(1-K-\varepsilon')} \varepsilon' + \eta.$$

Then, since  $\varepsilon' = O(\varepsilon)$ , it follows that

$$(5.15) \quad \varepsilon'' \approx \frac{1}{1-K} \varepsilon + \eta.$$

Then the error estimate (5.12) can be written approximately as follows:

$$(5.16) \quad |\tilde{x}_n^* - \bar{x}| \leq \frac{1}{1-K} \varepsilon + \eta.$$

This is the desired error estimate for  $\tilde{x}_n^*$  in the state of ONC.

Now, we consider the case where the computation is stopped by the criterion of the form

$$(5.17) \quad |\tilde{x}_{n+1}^* - \tilde{x}_n^*| \leq \tilde{\alpha}.$$

Let us put

$$M_2 = |\varphi(\tilde{x}_n^*) - \tilde{x}_n^*|,$$

then, from (5.8) and (5.9), it holds that

$$(5.18) \quad |\Psi^*(\tilde{x}_n^*) - \Psi(\tilde{x}_n^*)| \leq \varepsilon''',$$

where

$$(5.19) \quad \varepsilon''' = \frac{1}{1-K-\varepsilon'} \varepsilon + \frac{M_2}{(1-K)(1-K-\varepsilon')} \varepsilon' + \eta.$$

Hence, by the theory of Urabe [7], it is seen that

$$(5.20) \quad \begin{aligned} |\tilde{x}_{n+1}^* - \bar{x}| &\leq \frac{1 - \sqrt{1 - \frac{4L(\tilde{\alpha} + \varepsilon''')}{1-K}}}{1 + \sqrt{1 - \frac{4L(\tilde{\alpha} + \varepsilon''')}{1-K}}} \tilde{\alpha} \\ &\quad + \frac{2}{1 + \sqrt{1 - \frac{4L(\tilde{\alpha} + \varepsilon''')}{1-K}}} \varepsilon''' \\ &\approx \frac{L}{1-K} \tilde{\alpha}^2 + \varepsilon''' \end{aligned}$$

for the  $\tilde{x}_n^*$  satisfying the criterion (5.17). In this case, from (5.8) readily follows

$$|\varphi(\tilde{x}_n^*) - \tilde{x}_n^*| \leq (1+K)(\tilde{\alpha} + \eta) + \varepsilon.$$

Then, since  $\tilde{\alpha}, \eta \ll 1$  and  $\varepsilon' = O(\varepsilon)$ , we have

$$(5.21) \quad \varepsilon''' = \frac{\varepsilon}{1-K} + \eta.$$

Therefore we see that the error estimate (5.20) can be written approximately as follows:

$$(5.22) \quad |\tilde{x}_{n+1}^* - \bar{x}| \leq \frac{L}{1-K} \tilde{\alpha}^2 + \frac{1}{1-K} \varepsilon + \eta.$$

*This is the desired error estimate for  $\tilde{x}_{n+1}^*$  satisfying (5.17).*

When

$$|\varphi^*(\tilde{x}_n^*) - \tilde{x}_n^*| \leq \alpha,$$

it is readily seen from (5.8) that

$$|\tilde{x}_{n+1}^* - \tilde{x}_n^*| \leq \frac{\alpha}{1-K} + \eta.$$

Then, replacing  $\tilde{\alpha}$  by

$$\frac{\alpha}{1-K} + \eta$$

in (5.22), we have

$$|\tilde{x}_{n+1}^* - \bar{x}| \leq \frac{L}{1-K} \left( \frac{\alpha}{1-K} + \eta \right)^2 + \frac{1}{1-K} \varepsilon + \eta$$

$$\approx \frac{L}{(1-K)^3} \alpha^2 + \frac{1}{1-K} \varepsilon + \eta,$$

which is of the same form as (5.5). This fact means (5.22) implies (5.5), or, in other words, the error estimate (5.22) may be a little more precise than the error estimate (5.5).

Just like the AI-process relative to the AP-process, the MAI-process yields an acceleration of the MAP-process, consequently the former yields an acceleration of the initial CI-process much more than the latter.

**Example 2.** The MAI-process applied for the example of §3. Table 6 shows the results of the actual computation.

Table 6.

$n$	$\bar{x}_n^*$	$\bar{x}_{n+1}^* - \bar{x}_n^*$	$10^8 \times \tilde{e}_n$	$10^8 \times e_n$
0	0.2900 0000	-0.0070 6897		
1	0.2829 3103	-0.0000 8829	17850	8832
2	0.2828 4270	0.0000 0000	6	-1
3	0.2828 4270		3	-1

Here  $\tilde{e}_n$  refer to the error estimates given by (5.22) and  $e_n$  refer to the true errors. Table 6 shows that the error estimates given by (5.22) are considerably good and that the accurate values are obtained after a very few numbers of repetition of the iterative process

**Remark.** As is remarked in §§3-4, in using Aitken's processes, we have to stop the processes before  $|\varphi^*(x_n^*) - x_n^*|$  or  $|\varphi^*(\tilde{x}_n^*) - \tilde{x}_n^*|$  becomes so small that it may be  $O(\varepsilon)$ , for, otherwise, there may arise the overflow in the computer or may appear the indefinite form. On the contrary, *in using the modified Aitken's processes, there is not any restriction.* In addition, the results obtained by both processes have the accuracy of the same degree. Thus, if we can avail ourselves of the derivative of the given function  $\varphi(x)$ , the modified processes are preferable to the original Aitken's processes. But it is needless to say that the original Aitken's processes are preferable when the computation of the derivative of  $\varphi(x)$  is not convenient, for instance, when the analytical form of  $\varphi(x)$  is not given explicitly.

## 6. The actual SAP- and SAI-processes.

The simplified Aitken's processes are based on the function  $\Phi(x)$  defined by (1.14). Therefore

$$(6.1) \quad \Phi'(x) = (1+k) - k\varphi'(x).$$

Since  $k$  is chosen so that (1.15) may hold, we suppose

$$(6.2) \quad \max_{\xi \in [x', x'']} |(1+k) - k\varphi'(\xi)| \leq K(x', x'') \leq K_0 < < 1,$$

where  $[x', x''] \subset I$ .

Now the SAP-process is carried out in the actual computation as follows:

$$(6.3) \quad \bar{x}_{n+1}^* = x_n^* - k(x_{n+1}^* - x_n^*) + \eta_n,$$

where

$$x_{i+1}^* = \varphi^*(x_i^*) \quad (i = 0, 1, 2, \dots, n)$$

and  $\eta_n$  is a round-off error caused by the computation of  $k(x_{n+1}^* - x_n^*)$ . Therefore, from (6.3), it follows that

$$(6.4) \quad \begin{aligned} \bar{x}_{n+1}^* - \bar{x} &= (1+k)(x_n^* - \bar{x}) - k[\varphi(x_n^*) - \varphi(\bar{x})] - k\varepsilon_n + \eta_n \\ &= [(1+k) - k\varphi'(\xi)](x_n^* - \bar{x}) - k\varepsilon_n + \eta_n, \end{aligned}$$

where

$$\xi \in (\bar{x}, x_n^*).$$

Let us suppose

$$(6.5) \quad |(1+k) - k\varphi'(x_n^*)| \leq \kappa.$$

Then, from (2.6),

$$|(1+k) - k\varphi'(\xi)| \leq \kappa + |k|L \frac{\alpha + \varepsilon}{1-K}.$$

Consequently, from (6.4), we have

$$(6.6) \quad \begin{aligned} |\bar{x}_{n+1}^* - \bar{x}| &\leq \left( \kappa + |k|L \frac{\alpha + \varepsilon}{1-K} \right) \frac{\alpha + \varepsilon}{1-K} + |k|\varepsilon + |\eta_n| \\ &\approx \frac{1}{1-K} \kappa \alpha + \frac{|k|L}{(1-K)^2} \alpha^2 + |k|\varepsilon + |\eta_n|. \end{aligned}$$

*This is a desired error estimate for  $\bar{x}_{n+1}^*$  obtained by the SAP-process from  $x_n^*$  satisfying (2.7).*

**Example 1.** The SAP-process applied for the example of §3.

Let us take  $k$  so that

$$k = -3.5335 \approx \frac{x_{11}^* - x_{10}^*}{x_{12}^* - 2x_{11}^* + x_{10}^*}.$$

Then, for  $\varphi(x)$  under question,

$$(1+k) - k\varphi'(x_n^*) = 1 + kx_n^* = 1 - 3.5335x_n^*,$$

from which  $\kappa$  can be easily calculated. The computed results are shown in Table 7.

Table 7.

$n$	$\tilde{x}_n^*$	$x_{n+1}^* - x_n^*$	$\kappa$	$10^8 \times \bar{e}_n$	$10^8 \times e_n$
10		-0.0000 7166	0.0003 1954		
11	0.2828 4274	-0.0000 5138	0.0000 6633	34	3
12	0.2828 4274	-0.0000 3684	0.0001 1522	15	3
13	0.2828 4274	-0.0000 2641	0.0002 4540	11	3
14	0.2828 4275	-0.0000 1894	0.0003 3872	9	4
15	0.2828 4274	-0.0000 1358	0.0004 0564	7	3
16	0.2828 4274	-0.0000 0974	0.0004 5383	6	3
17	0.2828 4272	-0.0000 0699	0.0004 8804	5	1
18	0.2828 4270	-0.0000 0501	0.0005 1274	5	-1
19	0.2828 4271	-0.0000 0359	0.0005 3044	4	0
20	0.2828 4271	-0.0000 0258	0.0005 4313	4	0
21	0.2828 4269			4	-2

Here  $\bar{e}_n$  refer to the error estimates given by (6.6) and  $e_n$  refer to the true errors.

The SAI-process is carried out in the actual computation as follows:

$$\tilde{x}_{n+1}^* = \Phi^*(\tilde{x}_n^*) = \tilde{x}_n^* - k[\varphi^*(\tilde{x}_n^*) - \tilde{x}_n^*] + \eta_n,$$

where  $\eta_n$  is a round-off error caused by the computation of  $k[\varphi^*(\tilde{x}_n^*) - \tilde{x}_n^*]$ .

Therefore it is evident that

$$(6.7) \quad |\Phi^*(\tilde{x}_n^*) - \Phi(\tilde{x}_n^*)| \leq |k| \varepsilon + \eta,$$

where  $\eta$  is a number such that

$$(6.8) \quad |\eta_n| \leq \eta.$$

From (6.1) and (6.2), it is also evident that

$$(6.9) \quad |\Phi(x') - \Phi(x'')| \leq K(x', x'') |x' - x''|$$

for any  $x', x'' \in I$ .

Let us suppose

$$(6.10) \quad |(1+k) - k\varphi'(\tilde{x}_n^*)| \leq \tilde{\kappa},$$

then, from the definition of  $K(x', x'')$ , we may suppose

$$(6.11) \quad K(\bar{x}, \tilde{x}_n^*) = \tilde{\kappa} + |k|L|\tilde{x}_n^* - \bar{x}|.$$

By means of the method of Urabe [7], making use of (6.7) and (6.11), let

we derive the error estimates for the values obtained by the SAI-process.

First, for  $\tilde{x}_n^*$  in the state of ONC, let us derive the error estimates. By §2, let us put

$$\begin{cases} K_{p+1} = \tilde{\kappa} + |k|L\delta_p, \\ \delta_p = \frac{|k|\varepsilon + \eta}{1 - K_p} \quad (p = 0, 1, 2, \dots). \end{cases}$$

Then, in the limit where  $p \rightarrow \infty$ , we have

$$\hat{\delta} = \frac{|k|\varepsilon + \eta}{1 - \tilde{\kappa} - |k|L\hat{\delta}}$$

for  $\hat{\delta} = \lim_{p \rightarrow \infty} \delta_p$ . Therefore  $\hat{\delta}$  can be obtained by solving the Quadratic equation

$$\frac{|k|L}{1 - \tilde{\kappa}} \hat{\delta}^2 - \hat{\delta} + \frac{|k|\varepsilon + \eta}{1 - \tilde{\kappa}} = 0.$$

Since  $\hat{\delta} = 0$  for  $|k|\varepsilon + \eta = 0$ , we see that

$$(6.12) \quad \hat{\delta} = \frac{1 - \tilde{\kappa}}{2|k|L} \left[ 1 - \sqrt{1 - \frac{4|k|L}{(1 - \tilde{\kappa})^2} (|k|\varepsilon + \eta)} \right] \approx \frac{|k|\varepsilon + \eta}{1 - \tilde{\kappa}}.$$

Since  $|\tilde{x}_n^* - \bar{x}| \leq \hat{\delta}$  by the results of Urabe [7], the  $\hat{\delta}$  given by (6.12) provides the error bound for  $\tilde{x}_n^*$  in the state of ONC.

Next, for  $\tilde{x}_{n+1}^*$  such that

$$(6.13) \quad |\tilde{x}_{n+1}^* - \tilde{x}_n^*| \leq \tilde{\alpha},$$

let us derive the error estimates. By §2, let us put

$$\begin{cases} K_{p+1} = \tilde{\kappa} + |k|L\delta_p, \\ \delta_p = \frac{\tilde{\alpha} + |k|\varepsilon + \eta}{1 - K_p} \quad (p = 0, 1, 2, \dots). \end{cases}$$

Then, in the limit where  $p \rightarrow \infty$ , we have

$$\hat{K} = \tilde{\kappa} + |k|L \frac{\tilde{\alpha} + |k|\varepsilon + \eta}{1 - \hat{K}}$$

for  $\hat{K} = \lim_{p \rightarrow \infty} K_p$ . Therefore  $\hat{K}$  can be obtained by solving the Quadratic equation

$$\frac{1}{1 + \tilde{\kappa}} \hat{K}^2 - \hat{K} + \frac{\tilde{\kappa} + |k|L(\tilde{\alpha} + |k|\varepsilon + \eta)}{1 + \tilde{\kappa}} = 0.$$

Since  $\hat{K} = 0$  for  $\tilde{\kappa} = \tilde{\alpha} = \varepsilon = \eta = 0$ , we see that

$$\begin{aligned}
 (6.14) \quad \hat{K} &= \frac{1+\tilde{\kappa}}{2} \left[ 1 - \sqrt{1 - \frac{4\tilde{\kappa}'}{(1+\tilde{\kappa})^2}} \right] \\
 &= \frac{\tilde{\kappa}'}{1+\tilde{\kappa}} + \frac{\tilde{\kappa}'^2}{(1+\tilde{\kappa})^3} + \dots,
 \end{aligned}$$

where

$$(6.15) \quad \tilde{\kappa}' = \tilde{\kappa} + |k| L(\tilde{\alpha} + |k| \varepsilon + \eta).$$

Then, by the results of Urabe [7],

$$\begin{aligned}
 (6.16) \quad |\tilde{x}_{n+1}^* - \bar{x}| &\leq \frac{\hat{K}}{1-\hat{K}} \tilde{\alpha} + \frac{1}{1-\hat{K}} (|k| \varepsilon + \eta) \\
 &= \left[ \frac{1}{1+\tilde{\kappa}} \tilde{\kappa}' + \frac{(2+\tilde{\kappa})}{(1+\tilde{\kappa})^3} \tilde{\kappa}'^2 + \dots \right] \tilde{\alpha} + |k| \varepsilon + \eta + \dots \\
 &\approx \frac{1}{1+\tilde{\kappa}} \tilde{\kappa} \tilde{\alpha} + \frac{|k|L}{1+\tilde{\kappa}} \tilde{\alpha}^2 + \frac{2+\tilde{\kappa}}{(1+\tilde{\kappa})^3} \tilde{\kappa}^2 \tilde{\alpha} + |k| \varepsilon + \eta.
 \end{aligned}$$

This is the desired error estimate for  $\tilde{x}_{n+1}^*$  satisfying (6.13).

**Example 2.** The SAP-process applied for the example of §3. For  $\varphi(x)$  under question,

$$(1+k) - k\varphi'(x) = 1 + kx,$$

consequently let us take  $k$  so that

$$k = -3.45 \approx -\frac{1}{x_0^*} = -\frac{1}{0.29}.$$

The computed results are shown in Table 8.

Table 8.

$n$	$\tilde{x}_n^*$	$\tilde{x}_{n+1}^* - \tilde{x}_n^*$	$10^8 \times \tilde{\varepsilon}_n$	$10^8 \times \varepsilon_n$
0	0.2900 0000	-0.0070 7250		
1	0.2829 2750	-0.0000 8277	17605	8479
2	0.2828 4473	-0.0000 0197	208	202
3	0.2828 4276	-0.0000 0007	8	3
4	0.2828 4269	0.0000 0000	3	-2
5	0.2828 4269		3	-2

Here  $\tilde{\varepsilon}_n$  refer to the error estimates given by (6.16) and  $\varepsilon_n$  refer to the true errors.

**Remark.** Comparing (6.6) and (6.16) with (5.5) and (5.22), we see that,

*in the acceleration of the initial CI-process, the simplified Aitken's processes are both inferior to the modified ones as is predicted from their derivation. However, in the simplicity of the computation, the former is evidently superior to the latter.* So, in the actual computation, either of these processes should be selected according to whether accuracy or simplicity may be preferable.

However, when  $k$  is chosen so that  $\kappa, \tilde{\kappa} \ll 1$ , the inferiority of the simplified Aitken's processes in the acceleration is very slight as is seen from comparison of (6.6) and (6.16) with (5.5) and (5.22). So, in such a case, it is needless to say that the simplified Aitken's processes are preferable to the modified ones.

In conclusion, the author wishes to express his hearty gratitude to Prof. Urabe for his kind guidance and constant advice.

### References

- [1] Aitken, A. C., Studies in practical mathematics, 6: On the factorization of polynomials by iterative methods, Proc. Roy. Soc. Edinburgh. Sect. A, **16** (1837), 206-214.
- [2] Collatz, L., Einige Anwendungen funktionalanalytischer Methoden in praktischen Analysis, Z. Angew. Math. Phys., **4** (1953), 327-357.
- [3] Lance, G. N., Numerical Methods for High Speed Computers, London, 1960, pp. 151-152.
- [4] Ostrowski, A. M., Solution of Equations and Systems of Equations, New York and London, 1960, pp. 148-153.
- [5] Todd, J., A Survey of Numerical Analysis, New York, San Francisco, Toronto and London, 1962, pp. 5-6 and pp. 260-261.
- [6] Urabe, M., Convergence of Numerical Iteration in Solution of Equations, J. Sci. Hiroshima Univ. Ser. A, **19** (1956), pp. 479-489.
- [7] Urabe, M., Error Estimation in Numerical Solution of Equations by Iteration Process, J. Sci. Hiroshima Univ. Ser. A-I, **26** (1962), pp. 77-91.