Ivo Babuška

Problems of optimization and numerical stability in computations

Aplikace matematiky, Vol. 13 (1968), No. 1, 3–26

Persistent URL: http://dml.cz/dmlcz/103135

PROBLEMS OF OPTIMIZATION AND NUMERICAL STABILITY IN COMPUTATIONS[1])

IVO BABUŠKA

1. INTRODUCTION

Computer Science is a new scientific discipline. An important part of this discipline is the numerical mathematics. The "Art of Computation" is becoming a science; new questions and problems become important.

A typical problem is the problem of the creation of numerical methods, the determination of their "worth" and, in general, the choice of the most suitable method for the given purpose.

For example, the program-library in a computing centre contains mostly many algorithms for solving single mathematical problems. Opinions on the expedience of these algorithms are usually quite different and subjective. This statement is still more apparent when a method of applied mathematics is to be appreciated, especially in the field of scientific-technical computations. These scientific-technical computations are that part of the computer science in which I have some experience.

My paper will deal with questions which are more or less associated with this kind of computations.

I think that these computations may be characterized as a mathematical and constructive way of processing (transformation) of the given information to the required one[2]). I am sure that in scientific-technical computations it is necessary to emphasize the knowledge of information which we may collect and the appreciation of its reliability. Further it is necessary to formulate clearly the required information on the given problem. The necessity of a mathematical and constructive way of this processing is obvious here.

The clarity of the given and required information is an important part for a successful solution of a technical problem. Numerical mathematics are the rudiments of this constructive processing of information.

Numerical method generates (in a constructive manner) a mapping, from the class (space) of the given information to the class of the required one. It is important

[1]) In this paper some results obtained recently in Prague will be given.

[2]) HENRICI [23] defines numerical analysis as the theory of constructive methods in mathematical analysis (with emphasis on the word *constructive*).

that this mapping is defined on the entire *class* of information. This class will be the domain of definition of the given method (mapping).

Numerical process is an exact constructive law (prescription) of creation of the given mapping.

Computation is a concrete realization of the numerical process in the given case. We shall talk about *exact* realization when we compute without round-off errors and about a realization (or disturbed realisation) in a real computation.

It is obvious that there are many different manners of a constructive creation of one given mapping, i.e. many processes exist which transform the given information to the requested one and solve the same mathematical problem. It is evident that the question of choosing a process is very important.

It is clear that the choice and every optimization must necessarily be *relative* to the given information. This does not mean, however, that some methods might not be advantageous in a certain generality.

The manner in which we appreciate the method is of great importance. My experience is that, from the practical point of view, it is very important to respect an *incredulity* of the given information. This incredulity can be of different kinds. Some of them will be shown in the next part of the paper. It is essential that the method (and in general all conclusions) be stable with respect to these incredulities. I think that this stability is one of the most important points when choosing a method in practice.

In the next part I shall point out some aspects of these questions.

2. THE PROBLEM OF QUADRATURE FORMULAS[3]

In this section I shall show some aspects of ideas, which I mentioned previously, in a simple case of quadrature formulas.

Let our task be to determine numerically

$$J(f) = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,\mathrm{d}x\,. \tag{2.1}$$

We shall suppose that we know the following about the integrated function $f(x)$:

1. The function $f(x)$ is a continuous periodic function with the period 2π.
2. We can evaluate only the function $f(x)$ (i.e. compute the values of $f(x)$).

In this case, the simplest quadrature formula $T_n(f)$ is mostly used in practice, with

$$T_n(f) = \frac{1}{n}\sum_{j=1}^{n} f\left(\frac{2\pi}{n}j\right). \tag{2.2}$$

This formula is the well known trapezoid formula.

[3] In this part we are not dealing with the problems of the round-off error.

I will now analyse the question, if there are any reasons for selecting the trapezoid formula; we may ask e.g. why the Simpson-formula isn't better than the formula previously mentioned. Some arguments for choosing the trapezoid formula (in this case of integration of a periodic function) are included in some papers, e.g. MILNE [25], DAVIS [18] and others.

The error bounds for the trapezoid formula are studied in many papers. See [4], [5], [21], [24] and others. We will now analyze the problem of the choice of the quadrature formula according to the information we mentioned previously. In our considerations we shall confine the class of possible formulae to the linear one.

The choice of the quadrature formula means, in our case, to determine the sequence of linear functionals I_n in the form

$$I_n(f) = \sum_{k=1}^{n} a_{k,n} f(x_{k,n}), \quad 0 < x_{k,n} \leqq 2\pi \tag{2.3}$$

with the requirement that $I_n(f) \to J(f)$ for all functions $f(x)$ of the given class of functions.

We shall measure the amount of work in using a formula by the number of evaluations of the integrated function.

Let us now assume that $\boldsymbol{B}$ is a Banach space. Then we can define

$$\omega(n, \boldsymbol{B}) = \inf_{\substack{a_j, y_j \\ j=1,\dots,n}} \sup_{\|f\|_{\boldsymbol{B}} \leqq 1} \Big| \sum_{j=1}^{n} a_j f(y_j) - J(f) \Big| \tag{2.4}$$

and

$$\varrho(n, \boldsymbol{B}) = \inf_{a_j} \sup_{\|f\|_{\boldsymbol{B}} \leqq 1} \left| \sum_{j=1}^{n} a_j f\left(\frac{2\pi}{n} j\right) - J(f) \right| . \tag{2.5}$$

$\omega(n, \boldsymbol{B})$ is the minimal possible error under the assumption that we know only that $\|f\|_{\boldsymbol{B}} \leqq 1$. $\varrho(n, \boldsymbol{B})$ has an analogous meaning when we confine ourselves to use equidistant points in the quadrature formula. We shall further introduce

$$\Lambda(n, \boldsymbol{B}) = \sup_{\|f\|_{\boldsymbol{B}} \leqq 1} |T_n(f) - J(f)| . \tag{2.6}$$

$\Lambda(n, \boldsymbol{B})$ is evidently the error-bound of the trapezoid formula in the space $\boldsymbol{B}$. An objective measure of convenience of the given formula is given here by the comparison of $\Lambda(n, \boldsymbol{B})$ with $\omega(n, \boldsymbol{B})$, $\varrho(n, \boldsymbol{B})$ resp. This appreciation is obviously relative to the space $\boldsymbol{B}$.

The choice of the space $\boldsymbol{B}$ is very problematic in practice. In majority of cases there is a large incredulity as to whether it is convenient to take the integrated function as an element of a certain space $\boldsymbol{B}$. If the conclusion on the suitability of a formula is strongly dependent on the choice of $\boldsymbol{B}$, then the conclusion is not "stable" and it is not advantageous to use that formula in practice. Further we shall see that this "unstability" will appear in the case of the *optimal formula*, i.e. when we use the formula whose error equals $\varrho(n, \mathbf{B})$ or $\omega(n, \mathbf{B})$, then the results will strongly depend on the space $\mathbf{B}$. Conversely, a formula will be advantageous in practice if its error is *nearly* equal to $\varrho(n, \mathbf{B})$ or $\omega(n, \mathbf{B})$ but more or less *independent* of the space $\mathbf{B}$.

Later we shall see that only the trapezoid formula which is not an optimal one, has this property. We now introduce a class of Banach spaces of periodic functions.

Definition 2.1. *The Hilbert space $\mathbf{H}$ (over complex numbers) will be said to be periodic if:*

1. *Every $f \in \mathbf{H}$ is a 2π periodic, continuous function.*
2. *Let $\|f\|_c$ signify the norm in the space $\mathbf{C}$; then*

$$\|f\|_c \leqq \mathbf{C}(H)\,\|f\|_{\mathbf{H}}\,. \tag{2.7}$$

3. *If $f \in \mathbf{H}$, then $g(x) = f(x + c) \in \mathbf{H}$ for every real c and $\|f\|_{\mathbf{H}} = \|g\|_{\mathbf{H}}$.*

The space $\mathbf{H}$ will be said to be strongly periodic if it is periodic and if:

4. *$e^{ikx} \in \mathbf{H}$, $k = \ldots, -1, 0, 1, \ldots$ and $\|e^{ikx}\|_{\mathbf{H}} = \|e^{-ikx}\|_{\mathbf{H}}$.*
5. *If $|j| \geqq |k|$, then $\|e^{ijx}\|_{\mathbf{H}} \geqq \|e^{ikx}\|_{\mathbf{H}}$.*
6.

$$\|e^{i[n\alpha]x}\|_{\mathbf{H}}^2 \sum_{t=0}^{\infty} \|e^{i([n\alpha]+tn)x}\|_{\mathbf{H}}^{-2} \leqq D\,, \tag{2.8}$$

for $0 \leqq \alpha \leqq 2$, and D does not depend on n.

At the beginning of this section it was said that $f(x)$ is a periodic function. It is obvious that this information is insufficient. However, I think it is convenient to assume that the function $f(x)$ is an element of a periodic or strongly periodic space $\mathbf{H}$.

It is evident that now too we have a large incredulity as regards the concrete selection of the space $\mathbf{H}$. The importance of this incredulity is well seen in the next theorem and example.

Theorem 2.1. *Let $\mathbf{H}$ be a strongly periodic space with the norm*

$$\|f\|^2 = \int_0^{2\pi} (|f|^2 + A|f'|^2)\,\mathrm{d}x\,, \quad A > 0\,. \tag{2.9}$$

Then the error-bound of the formula

$$R_n^{(A)}(f) = C(n, \mathbf{H})\,T_n(f) \tag{2.10}$$

where

$$C^{-1}(n, \mathbf{H}) = 1 + \frac{2}{n^2} \sum_{t=1}^{\infty} \frac{1}{(tA)^2 + (1/n^2)}$$

is equal to $\varrho(n, \mathbf{H})$.

The theorem 2.1 affirms that the formula (2.10) is an optimal one if we are using the equidistant net. Now we shall introduce the following example:

Example 2.1. Let $f(x) = e^{\alpha \sin x}$, $\alpha = 3;10$. Then

$$J(f) = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,\mathrm{d}x = 4{\cdot}88079258586502408\ldots \text{ resp. } 2815{\cdot}71662846625447\,.$$

Table 2.1

Number of points n	$T_n(f)$, $f = e^{\alpha \sin x}$		$R_n^{(1)}(f)$, $f = e^{\alpha \sin x}$	
	$\alpha = 3$	$\alpha = 10$	$\alpha = 3$	$\alpha = 10$
8	4·88241999058958100	3047·90959481962441	4·64604604...	2900·35030...
16	4·88079258593666173	2815·77672896656761	4·81902223....	2780·14081...
24	4·88079258586502408	2815·71662897903758	4·85310536...	2799·74394...
Exact value	4·88079258586502408	2815·71662846625447		

In Tab. 2.1 we show the result obtained by the trapezoid formula $T_n(f)$ and by the optimal formula $R_n^{(A)}$ for $A = 1$. From this table we see that an optimal formula used in an inconvenient space may give bad results. We see that the conclusion of the convenience of the optimal formula is very "unstable" (with respect to the choice of $\boldsymbol{H}$). From this table we also see that the trapezoid formula ($C = 1$) gives very good results; however, the following theorem is true:

Theorem 2.2. *For every periodic space* $\boldsymbol{H}$

$$A(n, \boldsymbol{H}) > \varrho(n, \boldsymbol{H})\,. \tag{2.11}$$

This theorem shows that the trapezoid formula cannot be optimal in a periodic space. Nevertheless this formula is very advantageous in practice. The explanation of this fact can be seen in the following statement:

Theorem 2.3. *Let* $\boldsymbol{H}$ *be a periodic space. Then*

$$\lim_{n\to\infty} \frac{A(n, \boldsymbol{H})}{\varrho(n, \boldsymbol{H})} = 1\,. \tag{2.12}$$

No other formula has the property that the left-handside of (2.12) *is bounded for all periodic spaces* (*except for a finite number of indices of* n).

This theorem shows that the efficiency of the trapezoid formula is roughly the same as that of an optimal formula. This statement is now“ stable” with respect to the choice of $\boldsymbol{H}$. The asymptotic optimality (in the sense of (2.12)) of the trapezoid formula is valid only when the equidistant net is used, i.e. when we compare $\Lambda(n, \boldsymbol{H})$ with $\varrho(n, \boldsymbol{H})$.

The situation becomes more complicated when we compare $\Lambda(n, \boldsymbol{H})$ with $\omega(n, \boldsymbol{H})$. What happens will be seen from the following theorems:

Theorem 2.4. *For every sequence* $\zeta_1, \zeta_2, \ldots, \zeta_i > 0$ *there exists a periodic space* $\boldsymbol{H}$ *such that*

$$\limsup_{n \to \infty} \frac{\Lambda(n, \boldsymbol{H})}{\omega(n, \boldsymbol{H})\, \zeta_n} = \infty . \tag{2.13}$$

Theorem 2.5. *Let* $\boldsymbol{H}$ *be a strongly periodic space. Then we have*

$$\limsup_{n \to \infty} \frac{\Lambda(n, \boldsymbol{H})}{\omega(n, \boldsymbol{H}) \sqrt{n}} < \infty . \tag{2.14}$$

Theorems 2.4 and 2.5 show that it is reasonable to demand the universal efficiency relative to the set of *strongly* periodic spaces. The necessity to confine ourselves to some kind of incredulity is natural. This is a general statement when dealing with every kind of incredulity. I think, however, that the previously mentioned theorems show well the role of the trapezoid and optimal formulas and the role of the incredulity.

We have been dealing with the analysis of the convenience of the trapezoid formula $T_n(f)$. Now let us mention the error of $T_n(f)$. Practically all the known error estimations are based on the choice of the space $\boldsymbol{H}$ (or on the more general space $\boldsymbol{B}$).

The choice may be carried out a priori, i.e. before the computation or a posteriori, i.e. the choice is made with respect to the results obtained during the computation.

The error estimate is then

$$\varepsilon_n(f, \boldsymbol{H}) = \Lambda(n, \boldsymbol{H})\, \|f\|_{\boldsymbol{H}} . \tag{2.15}$$

The norm $\|f\|_{\boldsymbol{H}}$ has to be estimated (a priori or a posteriori).

There are many papers dealing with the estimation of $\Lambda(n, \boldsymbol{H})$ for different spaces, e.g. [5], [2], [24], [6] and others. Many results have been gathered in special books. See e.g. [26], [34] and others. With a suitable choice of $\boldsymbol{H}$ (resp. $\boldsymbol{B}$) we may obtain the estimates in $\boldsymbol{C}_n$, the “derivative-free” estimates (22); for the estimation of $\omega(n, \boldsymbol{H})$ and $\varrho(n, \boldsymbol{H})$ see e.g. [5], [8], [21], [36]. Moore’s important and principal results [30], [31], [32] may also be understood as an a posteriori choice of $\boldsymbol{B}$ and an a posteriori estimation of the norm $\|f\|_{\boldsymbol{B}}$. This a posteriori choice is made here by the computer. I want to emphasize that a choice of an a priori given class of possible spaces is given here. The problem of the optimal choice of a space in connection with the error estimation will now be discussed.

Table 2.2

Number of points n	$f = e^{10\sin x}$		
	$T_n(f)$	$\lvert J(f) - T_n(f)\rvert$	$\eta_n(f)$
8	3047·909594819624415	232·192966353369944	232·3732719565787845
16	2815·776728966567611	0·0601005003131402	0·0601005003142606
24	2815·716628979037584	0·0000005127831140	0·0000005127831167
32	2815·716628466254842	0·0000000000003720	0·0000000000003743
	$f = e^{50\sin x}$		
8	0·6480887567505754520 +21	0·3548333783656418203 +21	0·5063081633888883995 +21
16	0·3384555456320188415 +21	0·0452001672473682524 +21	0·0452789214681998668 +21
24	0·2951999264551136014 +21	0·0019445480701799697 +21	0·0019445486761790910 +21
32	0·2932816292532110631 +21	0·0000262508682774314 +21	0·0000262508682775479 +21
40	0·2932554985285131181 +21	0·0000001201435794864 +21	0·0000001201435794864 +21
48	0·2932553785869333396 +21	0·0000000002019997072 +21	0·0000000002019997072 +21

Let $f(x)$ be a 2π-periodic continuous function and let $\emptyset \neq \varkappa(f) = E(\boldsymbol{H};\ \boldsymbol{H}$ periodic, $f \in \boldsymbol{H})$.

Then the following theorem is valid:

Theorem 2.6. *Let*

$$f(x) = \sum_{k=-\infty}^{\infty} a_k e^{ikx} .$$

Then

$$\varepsilon_n(f) = \inf_{\boldsymbol{H} \in \varkappa(f)} \Lambda(n, \boldsymbol{H}) \, \|f\|_{\boldsymbol{H}} = \sum_{\substack{t=-\infty \\ t \neq 0}}^{\infty} |a_{tn}| . \tag{2.16}$$

There is often a simple way of obtaining this "best" estimation.

We introduce

Definition 2.2. *The function* $g(x) = \sum_{k=-\infty}^{\infty} b_k e^{ikx}$ *will be called an overfunction to the function* $f(x) = \sum_{k=-\infty}^{\infty} a_k e^{ikx}$, *if* $b_k \geqq |a_k|$, $k = \ldots -1, 0, 1, \ldots$ *This will be denoted by* $g \succ f$.

The overfunction can often be simply constructed. E.g. $e^{\sin x} \sin^2 x \prec \frac{1}{2} e^{\cos x} \,.\, (1 + \cos 2x)$. With this knowledge of the overfunction it is possible to get an estimation of $\varepsilon_n(f)$ in (2.16).

Theorem 2.7. *Let* $g \succ f$ *and* g''' *be a continuous function. Then*

$$\varepsilon_n(f) \leqq \eta_n(f) = \frac{1}{n^2} \left| T_n(g'') \right| . \tag{2.17}$$

The efficiency of this estimation will be shown in the following example:

Example 2.2. Let $f(x) = e^{\alpha \sin x}$, $\alpha > 0$. Then $f(x) \prec e^{\alpha \cos x}$. In table 2.2 we introduce the value $\eta_n(f)$ and $\sigma_n = |J(f) - T_n(f)|$ for $\alpha = 10; 50$. From table 2.2 a good accordance between the estimation and the real error may be seen. This error estimation is closely related to the ideas of Dahlquist [19] [20].

We dealt with the analysis of the computation of (2.1). Similar ideas can be used for the computation of the Fourier coefficients,

$$J_p(f) = \frac{1}{2\pi} \int_0^{2\pi} f(x) \, e^{ipx} \, \mathrm{d}x . \tag{2.18}$$

We obtain a principal new problem if we want to compute *simultaneously* the values J_{p_j} $j = 1, \ldots, k$. Obviously the simplest way is to compute these values *independently*. There is a question if it is possible to gain something when we make the computations simultaneously.

I will show it in the simplest case. Let us assume that we will compute both values J_0 and J_1 simultaneously. Put

$$(2.19) \qquad \Omega_{0,1}^{(n)}(\boldsymbol{H}) = \inf_{\substack{a_j, j=1,\dots,n \\ g(p), p=0,1}} \max_{p=0,1} \sup_{\|f\|_{\boldsymbol{H}} \leqq 1} \left| \sum_{j=1}^{n} a_j\, g(p) f\left(\frac{2\pi}{n} j\right) - J_p(f) \right|$$

$\Omega_{0.1}^{(n)}$ is apparently the minimal possible error in a simultaneous computation.

We shall analyse what can be gained by this kind of computation.

Let

$$(2.20) \qquad \pi_{0,1}^{(n)}(\boldsymbol{H}) = \max\left(A(n, \boldsymbol{H}), \|J_1\|_{\boldsymbol{H}}\right).$$

This is the error if we compute J_0 with the trapezoid formula and if we put $J_1(f) = 0$. The following theorem may be proved:

Theorem 2.8. *Let $\boldsymbol{H}$ be strongly periodic. Then*

$$\limsup_{n \to \infty} \frac{\pi_{0,1}^{(n)}(\boldsymbol{H})}{\Omega_{0,1}^{(n)}(\boldsymbol{H})} \leqq \sqrt{2}\,.$$

The theorem shows that we can gain practically nothing while performing a simultaneous computation. Theorem 2.8 is a special case of theorems which have been proved by P. Přikryl [33].

We analysed the case if only the function values were used in computing. All I said can be done if we use also the values of k derivatives. Here we shall assume besides (2.7) the following:

$$(2.7) \qquad \|f^{(s)}\|_{\boldsymbol{C}} \leqq C_s(\boldsymbol{H})\, \|f\|_{\boldsymbol{H}}\,, \quad s = 0, \dots, k$$

and

$$(2.8') \qquad \|e^{i[n\alpha]x}\|_{\boldsymbol{H}}^{2} \sum_{t=0}^{\infty}{}' \|e^{i([n\alpha]+tn)x}\|_{\boldsymbol{H}}^{-2} (t + \alpha)^{2k} \leqq D$$

$0 \leqq \alpha \leqq 2k + 2$ and D does not depend on n.

In this case the space $\boldsymbol{H}$ will be said k-periodic or k-strongly periodic.

Analogously to (2.5) we now have

$$(2.22) \qquad \varrho_k(n, \boldsymbol{B}) = \inf_{\substack{a_j^{(s)} \\ j=1,\dots,n \\ s=0,\dots,k}} \sup_{\|f\| \leqq 1} \left| \sum_{s=0}^{k} \sum_{j=1}^{n} a_j^{(s)} f^{(s)}\left(\frac{2\pi}{n} j\right) - J(f) \right|.$$

Now I shall mention a special result of Segeth (See [35]) who studied this field of problems. One of the problems here is, roughly speaking, the following:

Is it better to use more values of a functions in the quadrature or is it better to compute and use the values of the derivatives?

An answer to this is given by the comparison between ϱ_0 and ϱ_k. It can be shown that for 2-strongly periodic spaces

$$(2.23) \qquad \varrho_2(n, \boldsymbol{H}) = \inf_{\substack{a_j^{(s)}, j=1,\dots,n,\\ s=0,2}} \sup_{\|f\|_{\boldsymbol{H}} \leqq 1} \left| \sum_{\substack{s=0\\ s\neq 1}}^{2} \sum_{j=1}^{n} a_j^{(s)} f^{(s)}\left(\frac{2\pi}{n} j\right) - J(f) \right| .$$

Let us assume that the amount of work needed for the evaluation of $f(x)$ is equal to 1 and that for the derivative is α. Then the whole work with the use of n points will be $n(1 + \alpha)$. This values will be the measure of the "work" when using the given formula with n points.

$$(2.24) \qquad S(\alpha, \boldsymbol{H}) = \limsup_{n\to\infty} \frac{\varrho_2(n, \boldsymbol{H})}{\varrho([n(1 + \alpha)], \boldsymbol{H})}$$

gives now the required answer (relatively to the space $\boldsymbol{H}$).

Thus, for example, the following theorem is true:

Theorem 2.9. *Let $\boldsymbol{H}$ be a 2-strongly periodic space. Let $\alpha \geqq 1$. Let $\|e^{inx}\|_{\boldsymbol{H}}^2 = g(n^2)$ where g is an entire function. Then $S(\alpha, \boldsymbol{H}) > 1$. If g is not a polynomial then $S(\alpha, \boldsymbol{H}) = \infty$ for $\alpha > 1$ and $S(1, \boldsymbol{H}) = 3$.*

Theorem 2.9 shows more or less that if the amount of work needed for the evaluation of derivatives is not less than that needed for the evaluation of the function, it is not advantageous to use the formula with derivatives.

Previously in this section we dealt with the trapezoid formula T_n. An analogous role is played here by the formula

$$(2.25) \qquad T_n^{(2)}(f) = \frac{1}{n} \sum_{k=1}^{n} f\left(\frac{2\pi}{n} k\right) + \frac{1}{n^3} \sum_{k=1}^{n} f''\left(\frac{2\pi}{n} k\right).$$

There is also a theorem analogous to Theorem 2.9 for the use of (2.25), given more exactly and in more detail in (25). As an illustration I shall give the following example:

Example 2.3. Compute also $J(f) = 1/2\pi \int_0^{2\pi} f(x)\,\mathrm{d}x$ for $f(x) = e^{\beta \sin x}$, $\beta = 10;50$. Let us assume $\alpha = 1$. In table 2.3 we see the error when using the formulas T_n and $T_n^{(2)}$ in dependence on the amount of work (i.e. on n resp. $n(1 + \alpha)$).

We see that the computation without the use of derivatives is more advantageous. This agrees fully with the theoretic investigations. In accordance with the theorem the error of the formula with derivatives is nearly three times larger than that of the formula without derivatives. All we said was connected with the computation of (2.1) and (2.18) respectively.

Now I shall briefly speak about the computation of

$$(2.26) \qquad J_g(f) = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, g(x)\,\mathrm{d}x\,, \quad g \in L_2\,.$$

Table 2.3

Amount of work n	$f(x) = e^{10\sin x}$		$f(x) = e^{50\sin x}$	
	Error of the formula without derivatives (T_n)	Error of the formula with derivatives $(T_n^{(2)})$	Error of the formula without derivatives (T_n)	Error of the formula with derivatives $(T_n^{(2)})$
16	0·60100 -1	0·18030 0	0·45200 $+20$	0·15147 $+21$
32	0·37 -12	0·11200 -11	0·26250 $+17$	0·78754 $+17$
48	—	—	0·20199 $+12$	0·60599 $+12$

We shall not analyse all the problems associated with this computation. All can be done analogously. The formula which plays the same role here as the trapezoid formula is the following (see (6)):

$$T_n^{(g)}(f) = \frac{1}{n}\sum_{j=1}^{n} S_n\left(\frac{2\pi}{n}j\right) f\left(\frac{2\pi}{n}j\right) \tag{2.27}$$

where

$$S_n(x) = \sum_{k=-[n/2]+1}^{[n/2]} b_k^{(n)} e^{ikx}\,,$$

$$b_k^{(n)} = b_k \quad \text{for} \quad k < \left[\frac{n}{2}\right],$$

$$b_{n/2}^{(n)} = \tfrac{1}{2}(b_{n/2} + b_{-n/2})\,, \qquad \frac{n}{2} = \left[\frac{n}{2}\right].$$

$$g(x) = \sum_{k=-\infty}^{\infty} b_k e^{ikx}\,.$$

The error estimation by an everfunction can be made. As an illustration I shall show

Example 2.4. Compute

$$L = \int_{-\pi/2}^{+\pi/2} e^{\alpha \sin x} \cos x \,\mathrm{d}x\,, \quad \alpha = 1;5\,. \tag{2.28}$$

Apparently this integral may be written in the form (2.26). Obviously it can also be written like this:

$$L = \int_{-1}^{+1} e^{\alpha x} \,\mathrm{d}x\,. \tag{2.29}$$

In the table 2.4 we have shown the errors of (2.27) in comparison with the Romberg's integration (see [16], [17]). In the table there is also shown the error obtained by the use of the overfunction $e^{\cos x} \cos x$. The computation was made with computer ICT 1905 in double precision.

Table 2.4

Number of evaluations of function n	$f = e^{\sin x}$		$f = e^x$
	Error of (2.27) for (2.28)	Estimation based on overfunction	Error of Romberg formula of (2.29)
9	0·634 − 8	0·171 − 5	−0·421 −10
17	0·271 −19	0·245 −16	−0·416 −14
25	0	0·437 −18	—
33	0	—	0·407 −19
	$f = e^{5 \sin x}$		$f = e^{5x}$
9	0·510 − 2	0·381 0	−0·844 − 3
17	0·254 − 8	0·126 − 5	−0·208 − 5
25	0·492 −16	0·801 −13	—
33	−0·217 −18	0·268 −16	0·128 − 8

In a simple case of quadrature we have shown some aspects of incredulity with respect to the given information and the meaning of "stability" of a conclusion. It is possible to generalize these ideas in different ways. A possibility of a generalization can be seen in [28]. I shall, however, not deal with it here.

3. BOUNDARY-VALUE PROBLEMS FOR ORDINARY DIFFERENTIAL EQUATIONS

In section 2 we showed one kind of incredulity as regards the given information and how to deal with it. I shall now mention some ether aspects of incredulity. A simple problem will again be analysed. Let us solve the following boundary-value problem

$$(p(x)\, y')' - q(x)\, y = f(x) \tag{3.1}$$

with the boundary conditions

$$y(0) = a\,, \quad y(L) = b\,. \tag{3.2}$$

We assume that $p(x)$, $g(x)$ and $f(x)$ are sufficiently smooth and $p(x) > \alpha$, $q(x) \geqq 0$.

The functions p, q, f have a physical meaning. Nevertheless, we know them only approximately in practice.

Let the possible disturbances (incredulities) of p, q, f be σ, ϱ, φ respectively. From the physical point of view these perturbances are small *in a certain sense* (norm). They may also have further properties. Such perturbances will be called admissible disturbancies. We shall assume that small admissible disturbances result in a small change in the solution.

It is well known that a numerical process cannot be realized with an absolute exactness. Every realization of a process by computation is disturbed (by round-off errors). We can mostly imagine, however, this *disturbed* realization as an exact one (without disturbance) but with the disturbed given information. We shall speak about *replaced disturbances* (of information) in this case.[4] It is reasonable to speak about a *suitable numerical process* if the replaced disturbances

a) are admissible

b) the order of disturbances is the same as the order of error in the individual operations.

Bauer [13], [14], [15] used a similar approach in his investigations of numerical processes in algebraic problems.

There are *suitable* and *non suitable* processes. I shall show them by the process of solving (3.1) and (3.2).

Example 3.1. The method of combination of solutions leads to a non suitable process. This method, as known, consists in solving two initial-value problems for the initial conditions $y(0) = 0$, $y(0) = \gamma_j$, $j = 1, 2$ and the required solution (3.1), (3.2) is determined by a suitable combination. Let $p(x) = (1 + x)$, $q(x) = 500$, $f(x) = \pi \cos \pi x - (500 + \pi^2(1 + x)) \sin \pi x$, $L = 1$, $a = b = 0$. We solve the initial problem by the Runge-Kutta-Gill method of the 4th order for step $h = 0{,}025$ (computer LGP 30) See [12]. The results obtained are given in Table 3.1.

Example 3.2. The factorization method leads to a suitable process. By this method (see e.g. [12]) we solve the following system

$$\Psi' + q\Psi^2 = \frac{1}{p}, \quad \Psi(0) = 0,$$

$$u' + q\Psi u = f\Psi, \quad u(0) = -a,$$

$$\Psi y' - \frac{y}{p} = \frac{u}{p}, \quad y(L) = b.$$

Let us solve the same problem as in Example 3.1 by this method. The initial problems are also solved with Runge-Kutta-Gill method with $h = 0{\cdot}025$. We obtain the results mentioned in Table 3.1.

[4] The method of replaced distrubances (backward-method) was used with large success by WILKINSON. See [42], [43].

I have said that we can mostly consider the disturbed realization of a process as an exact realization with the disturbed input (i.e. given) information. In this case in the method of factorization the replaced disturbances are small in the following norms: σ, ϱ in $\mathbf{C}$ norm and φ in the norm $\|\varphi\| = \|\int_0^x \varphi \, \mathrm{d}x\|_{\mathbf{C}}$. It may be seen that these disturbances are admissible.

Table 3.1

x	$y(x)$ by method of combination	Exact solution	$y(x)$ by method of factorization
0·100	0·3090103	0·3090170	0·3090018
0·400	0·9510075	0·9510565	0·9510461
0·500	1·005031	1·0000000	0·9999897
0·700	0·8577343	0·8090170	0·8090081
0·750	1·374171	0·7071068	0·7070985
0·800	0·0000000	0·5877852	0·5877778
0·900	9·700032	0·3090170	0·3090119

It is obvious that the question of existence of a suitable numerical process for the solution of the given problem is very important. The method of factorization may be generalized to a general boundary (or multipoint) problem for the system

$$x'(s) - A(s)\,x(s) = f(s)\,. \tag{3.4}$$

J. Taufer (see [38], [39]) has investigated in detail the replaced disturbances for a concrete kind of factorization and has shown that his factorization method is convenient in the previously mentioned sense. Another kind of factorization method, sometimes called method of the transfer of boundary conditions, was investigated in recent years, for example, by Abramov (see [1], [2]) who also briefly mentioned the possibility of showing the suitability of this process for the general case (3.4). See [3].

In [7] and [12] the stability of the differential equations of the factorization method in special cases has been studied.

Example 3.3. As an example I shall show the computation of a continuous beam of 20 fields built-in at the end constantly loaded. In practice, the method of transfer of matrices which is very similar to the method of combination of solutions is very often used. See e.g. [45].

In the following table 3.2 there are shown the moments at some supports computed by the usual method as well as by Taufer's factorization method. The previously mentioned factorization method can also be used in solving the eigenvalue problem. See [40].

Table 3.2

Number of support	Exact moment at the support	Computed moment at the support by mentioned method	Computed moment at the support by factorization method
4	5·000000 −3	4·9999999 −3	4·999999999 −3
12	5·000000 −3	4·9492238 −3	5·000000006 −3
19	5·000000 −3	1·618765 −1	5·000000004 −3
20	5·000000 −3	7·790814 −1	4·999999999 −3

4. STABILITY OF NUMERICAL PROCEASES

In the previous sections we dealt with some aspects of incredulity as to the choice of a numerical process. In this section we shall deal with a quantitative characterization of the numerical stability of a given numerical process. See [9], [10], [11], [12]. In computations of problems of mathematical analysis, the existence of a subscript n (e.g. number of steps) is typical so that we obtain the required result only for $n \to \infty$. In section 1 we introduced a numerical process. Here we shall define it more exactly.

Definition 4.1. *Let there be given a sequence of normed vector spaces*

$$\mathbf{X}^{(n)}_{-p_n}, \mathbf{X}^{(n)}_{p_n+1}, \ldots, \mathbf{X}^{(n)}_0, \mathbf{X}^{(n)}_1, \ldots, \mathbf{X}^{(n)}_{N_n}, \quad n = 1, 2, \ldots$$

and a sequence of continuous operators

$$\mathbf{A}^{(n)}_i, \; i = 0, 1, \ldots, \quad N_n - 1, \; n = 1, 2, \ldots$$

mapping the Cartesian product

$$\mathbf{X}^{(n)}_{-p_n} \times \mathbf{X}_{-p_n+1} \times \ldots \times \mathbf{X}^{(n)}_i \quad \textit{into} \quad \mathbf{X}^{(n)}_{i+1},$$

Further let the sets

$$M^{(n)}_k \subset \mathbf{X}^{(n)}_k \quad \textit{for} \quad k = -p_n, \quad -p_n + 1, \ldots, 0$$

be given. Then the sequence of equations

$$x^{(n)}_{i+1} = \mathbf{A}^{(n)}_i \left(x^{(n)}_{-p_n}, x_{-p_n+1}, \ldots, x^{(n)}_i\right),$$

$$i = 0, 1, \ldots, N_n - 1, x^{(n)}_k \in \mathbf{X}^{(n)}_k, \quad k > 0,$$

$$x^{(n)}_k \in M^{(n)}_k, \quad k \leqq 0, \; n = 1, 2, \ldots$$

will be called a numerical process. The set $M^{(n)}_k$ will be called the set of input data and the elements $x^{(j)}_k$ $k = 1, 2, \ldots, N_n$ will be called the solution corresponding to input elements $x^{(n)}_k$, $k = -p_n, \ldots, 0$.

In practice, the numerical processes as by Definition 4.1 cannot be solved exactly by the computer (round-off errors). Hence we introduce the following definition:

Definition 4.2. *Let there be given a numerical proces in thesense of Definition 4.1. Let there be given the input elements* $x_k^{(n)}$, $k = -p_n, \ldots, 0$ *and a sequence of numbers* $\{a_j^{(n)}\} = \xi^{(n)}$, $a_j^{(n)} \geqq 0$, $j = -p_n, \ldots, N_n$; $n = 1, 2, \ldots$ *and denote* $\tilde{x}_i^{(n)} \in \mathbf{X}_k^{(n)} = -p_n, \ldots, 0, 1, \ldots, N_n$ *the elements satisfying the equations*

$$(4.2) \qquad \tilde{x}_{i+1}^{(n)} = \mathbf{A}_i^{(n)}\left(\tilde{x}_{-p_n}^{(n)}, \tilde{x}_{-p_n+1}^{(n)}, \ldots, \tilde{x}_i^{(n)}\right) + \vartheta_{i+1}^{(n)}, \quad i = 1, \ldots, N_n - 1 .$$

$$(4.3) \qquad \tilde{x}_i^{(n)} = x_i^{(n)} + \vartheta_i^{(n)}, \quad \tilde{x}_i^{(n)} \in M_i^{(n)}, \quad i = -p_n, \ldots, 0 .$$

The solution of the given numerical process corresponding to input elements $x_k^{(n)}$, $k = -p_n, \ldots, 0$ *and to the sequence* $\xi^{(n)}$ *will be called* β_s*-solution, if*

$$(4.4) \qquad \limsup_{\Delta \to 0} \frac{1}{\Delta} \sup_{|\vartheta_i(n)| \leqq a_i(n).\Delta,\ i = -p_n, \ldots, N_n} \sup \left|\tilde{x}_i^{(n)} - x_i^{(n)}\right| \leqq Cn^s$$

and C does not depend on n.

We will speak about B_{so}*-solution if* $S_0 = \inf s$.

The investigations of concrete given processes have been done in the previously mentioned way in many cases. See e.g. [12], [27], [32], [41], [44] and others.

I shall now give some examples explaining the meaning of the previous definitions. Let us solve the initial problem for an ordinary differential equation

$$(4.5) \qquad y' = f(x, y), \quad y(a) = y .$$

The Runge-Kutta method can be written as follows

$$y_{i+1}^{(n)} = y_i^{(n)} + h^{(n)} \Phi_f(x_i^{(n)}, y_i^{(n)}, h^{(n)}) .$$

A slight change will be made to simplify the notation in Definitions 4.1 and 4.2. We shall investigate two processes

$$\text{I.} \qquad y_{i+1}^{(n)} = y_i^{(n)} + h^{(n)} \Phi_f(x_i^{(n)}, y_i^{(n)}, h^{(n)}), \quad i = 0, 1, 2, \ldots$$

$$(4.6) \qquad y_0^{(n)} = y, \quad x_{i+1}^{(n)} = x_i^{(n)} + h^{(n)}, \quad i = 0, 1, 2, \ldots, x_0^{(n)} = a,$$

$$h^{(n)} = \frac{C}{n} .$$

$$\text{II.} \qquad y_{i+1}^{(n)} = y_i^{(n)} + h^{(n)} \Phi_f(x_i^{(n)}, y_i^{(n)}, h^{(n)}), \quad i = 0, 1, 2, \ldots$$

$$(4.7) \qquad y_0^{(n)} = y, \quad x_{i+1}^{(n)} = a + ih^{(n)}, \quad i = 0, 1, 2, \ldots$$

$$h^{(n)} = \frac{C}{n} .$$

Let further

$$(4.8) \qquad y_i^{(n)} \in \mathbf{Y}_i^{(n)}, \quad h^{(n)}\Phi_f \in \mathbf{Z}_i^{(n)}, \quad x_i^{(n)} \in \mathbf{X}_i^{(n)}, \quad h^{(n)} \in \mathbf{H}^{(n)} .$$

The spaces $\mathbf{X}_i^{(n)}, \mathbf{Y}_i^{(n)}, \mathbf{Z}_i^{(n)}, \mathbf{X}^{(n)}$ are spaces of real numbers with the norm $|x|$. The meaning of the previously mentioned mapping is evident in this case. The numerical process is also clear. The disturbed process is as follows

I.

$$\tilde{y}_{i+1}^{(n)} = \tilde{y}_i^{(n)} + \tilde{z}_i^{(n)} + {}^{\mathrm{I}}\vartheta_{i+1}^{(\mathbf{Y},n)}, \tag{4.9}$$

$$\tilde{x}_{i+1}^{(n)} = \tilde{x}_i^{(n)} + \tilde{h}^{(n)} + {}^{\mathrm{I}}\vartheta_{i+1}^{(\mathbf{X},n)}, \tag{4.10}$$

$$\tilde{z}_i^{(n)} = \tilde{h}^{(n)}\,\Phi(\tilde{x}_i^{(n)}, \tilde{y}_i^{(n)}, \tilde{h}^{(n)}) + {}^{\mathrm{I}}\vartheta_i^{(\mathbf{Z},n)}, \tag{4.11}$$

$$\tilde{h}^{(n)} = \frac{C}{n} + {}^{\mathrm{I}}\vartheta^{(\mathbf{H},n)}. \tag{4.12}$$

II. The equations (4.9), (4.11), (4.12) remain unchanged. (4.10) now has the following form

$$\tilde{x}_{i+1}^{(n)} = a + ih^{(n)} + {}^{\mathrm{II}}\vartheta_{i+1}^{(\mathbf{X},n)}. \tag{4.10'}$$

We can compute these processes in a different manner. These computations differ in disturbances. The following mathematical models can be assumed

a) Fixed point computation

$$\left|{}^{j}\vartheta_i^{(\mathbf{Y},n)}\right| \leqq \Delta, \quad \left|{}^{j}\vartheta_i^{(\mathbf{X},n)}\right| \leqq \Delta,$$
$$\left|{}^{j}\vartheta_i^{(\mathbf{Z},n)}\right| \leqq \Delta, \quad \left|{}^{j}\vartheta^{(\mathbf{H},n)}\right| \leqq \Delta, \quad j = \mathrm{I}, \mathrm{II}.$$

b)

$$\left|{}^{j}\vartheta_i^{(\mathbf{Y},n)}\right| \leqq \Delta, \quad \left|{}^{j}{}_i^{(\mathbf{X},n)}\right| \leqq \Delta,$$
$$\left|{}^{j}\vartheta_i^{(\mathbf{Z},n)}\right| \leqq h^{(n)}\Delta, \quad \left|{}^{j}\vartheta^{(\mathbf{H},n)}\right| \leqq h^{(n)}\Delta, \quad j = \mathrm{I}, \mathrm{II}.$$

b′) Normalized floating point computation

$$\left|{}^{j}\vartheta_i^{(\mathbf{Y},n)}\right| \leqq \Delta\left|\tilde{y}_i^{(n)}\right|, \quad \left|{}^{j}\vartheta_i^{(\mathbf{X},n)}\right| \leqq \Delta\left|\tilde{x}_i^{(n)}\right|,$$
$$\left|{}^{j}\vartheta_i^{(\mathbf{Z},n)}\right| \leqq \Delta\left|\tilde{z}_i^{(n)}\right|, \quad \left|{}^{j}\vartheta^{(\mathbf{H},n)}\right| \leqq \Delta\left|\tilde{h}^{(n)}\right|, \quad j = \mathrm{I}, \mathrm{II}.$$

c) Normalized floating point computation with computation (4.9) in the process II in double precision

$$\left|{}^{\mathrm{II}}\vartheta_i^{(\mathbf{Y},n)}\right| = 0, \quad \left|{}^{\mathrm{II}}\vartheta_i^{(\mathbf{X},n)}\right| \leqq \Delta\left|\tilde{x}_i^{(n)}\right|, \tag{4.13}$$
$$\left|{}^{\mathrm{II}}\vartheta_i^{(\mathbf{Z},n)}\right| \leqq \tilde{z}_i^{(n)}\right| \Delta, \quad \left|{}^{\mathrm{II}}\vartheta^{(\mathbf{H},n)}\right| \leqq \left|\tilde{h}^{(n)}\right| \Delta.\ {}^{5})$$

The sequence $\xi^{(n)}$ is obvious and I shall not describe it. The following theorem may be proved.

[5]) The same effect can be obtained as follows: $\bar{y}_{i+1}^{(n)} = \bar{y}_i^{(n)} + h^{(n)}\Phi_f{}'x_i^{(n)}.\ y_i^{(n)}, h^{(n)}),\ \eta_{i+1}^{(n)} = (\bar{y}_{i+1}^{(n)} - y_i^{(n)}) - h^{(n)}\Phi_f(x_i^{(n)}, y_i^{(n)}, h^{(n)}),\ \bar{y}_0^{(n)} = y,\ \varepsilon_{i+1}^{(n)} = \varepsilon_i^{(n)} - \eta_{i+1}^{(n)},\ \varepsilon_0^{(n)} = 0,\ y_i^{(n)} = \bar{y}_i^{(n)} + \varepsilon_i^{(n)}$.

Theorem 4.1. *The previously mentioned processes are*

a) B_1 *solution,*

b), b') B_1 *solution,*

c) B_0 *solution.*

I shall now show the meaning of this theorem by means of the following example:

Example 4.1. We shall solve the initial problem for the equation

$$y' = x(x + 2)\,y^3 + (x + 3)\,y^2\,, \tag{4.14}$$

$$y(a) = -\frac{2}{a(a + 2)}\,, \tag{4.15}$$

with the standard Runge-Kutta method of the 4th degree. $y(a + \frac{1}{2})$ is to be solved. Here we obviously have $C = \frac{1}{2}$. Our task is to estimate $\tilde{y}_n^{(n)} - y_n^{(n)}$ in dependence on n. Since the solution of (4.14) and (4.15) is $y(x) = -[2/x(x + 2)]$ and we do not know $y_i^{(n)}$, we shall use $y_n^{(n)} \to y(a + \frac{1}{2})$ and put $\varepsilon^{(n)} = |y_n^{(n)} - y(a + \frac{1}{2})|$. In the following figures there are the outcomes of computations. In fig. 4.1 a there are the results for the process I.b, $a = 0{,}5$, obtained with MINSK 22.[6]) The parameter n has been selected as a decadic value.

It is interesting to ask what happens if we investigate $\eta^{(n)} = |\tilde{y}_n^{(n)} - y(\tilde{x}_n^{(n)})|$. It may be shown that this is also a B_1 solution. In fig. 4.1b we see the results.

A further interesting question is what happens if we use n diadic. In fig. 4.2 we also see the results $\varepsilon^{(n)}$ for n diadic.

For a diadic n exactly the same results for computation I.b and II.b are obtained. In fig. 4.3 and 4.4 we have the results for I.b n decadic and II.b.

In fig. 4.5 we have the results for II.c and $a = 300$.

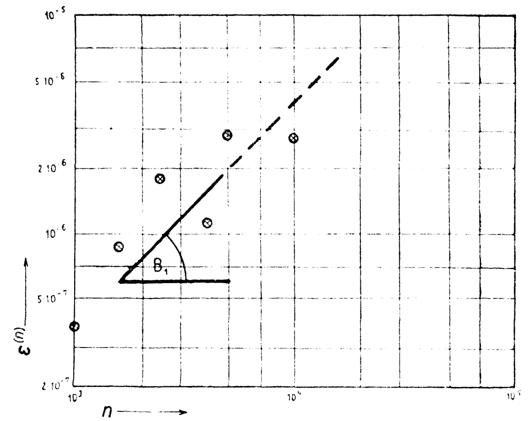
From the mentioned example we can clearly see that in computations there are different kinds of importance with respect to the stability. E.g. we have seen that the floating point makes the round-off error smaller but the results remain unchanged. Such a kind of considerations may be very valuable in practice, yet we cannot deal with it here.

It is obvious that the knowledge of stability, especially the B_s stability is an important factor in a suitable choice of method.

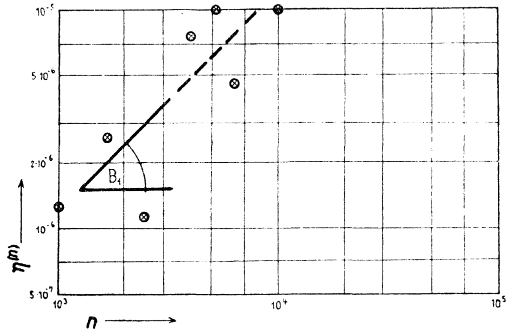
There is a question if it exist a B_0 solution for computation in simple precision normalized floating point. We have seen that the answer is possitive (see 5)).

I will show also the example of the quadrature formula T_n. Let $n = 2^k$. Let the

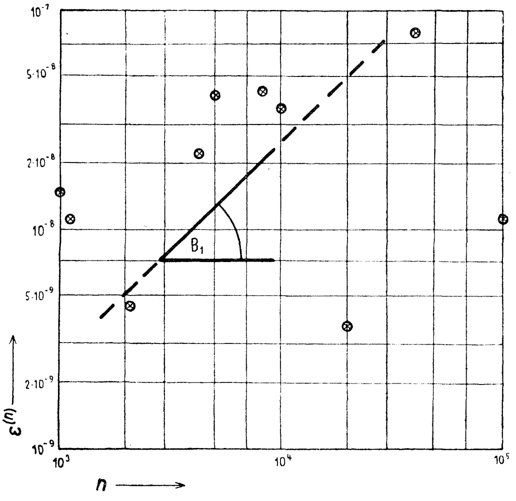
[6]) This computer is a diadic one.



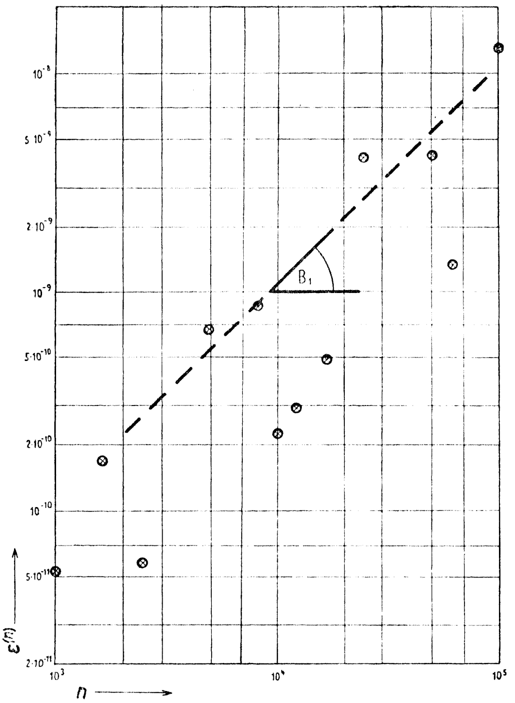
⟨0·5, 1⟩
I.b decadic
Fig. 4.1a



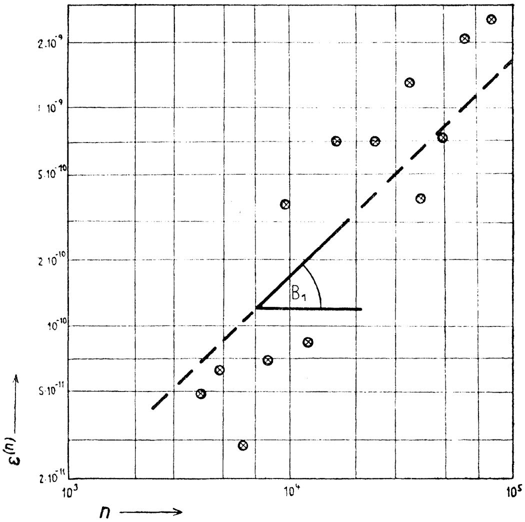
⟨0·5, 1⟩
I.b decadic
Fig. 4.1b



⟨0·5, 1⟩
II.b
Fig. 4.2



⟨300, 300·5⟩
I.b
Fig. 4.3



⟨300, 300·5⟩
II.b
Fig. 4.4

process of the computation of T_n be the following

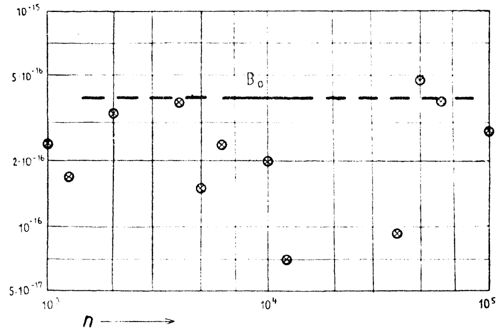
$$P_s = 2^{-s}\sum_{j=1}^{2^{s-1}}\left[2f\left(\frac{2\pi}{2^s}(2j-1)\right) - f\left(\frac{2\pi}{2^s}(2j-2)\right) - f\left(\frac{2\pi}{2^s}2j\right)\right], \quad s = 1, 2, \ldots, k.$$

$$P_0 = \tfrac{1}{2}(f(0) + f(2\pi)),$$

$$K_{n+1} = K_n + P_{k-n}, \quad K_0 = 0, \quad n = 0, 1, \ldots, k,$$

$$K_{k+1} = T_n.$$

⟨300, 300·5⟩
II.c
Fig. 4,5



Then the computation is a B_0 solution provided that we compute with normalized floating point and simple precision[7]).

We shall show an other interesting example. Let us solve the initial problem for the differential equation

$$y'' = f(x, y). \tag{4.8}$$

The usual difference method leads to the following formula

$$y_{n+3} - 2y_n + y_{n+1} = \tfrac{1}{12}h^2(13f_{n+2} - 2f_{n+1} + f_n). \tag{4.9}$$

This formula can be written in the following form

$$z_{n+1} - z_n = hf_n, \tag{4.10}$$

$$y_{n+2} - y_{n+1} = \tfrac{1}{12}(13z_{n+2} - 2z_{n+1} + z_n)$$

and the following theorem is true:

Theorem 4.2. *The numerical process based on* (4.9) *resp.* (4.10) *is a B_2, B_1 process respectively for $\xi^{(n)} = \{1, 1, \ldots\}$.*

This example (see [12], [41]) shows the possibilities of getting a better stability through simple changes in the method. The question when it is possible to write a formula in a form having a better stability is solved in [41].

[7]) It is possible evaluate the sum $S_n = \sum_{k=1}^{n} a_k$ with $n - 1$ operations so that the total round-off error ε_m is of the order $\lg_2 n \sum_{k=1}^{n} |a_k|$,

As a further example I shall show the numerical stability of the numerical process of overrelaxation for usual finite-difference equations for second-order elliptic partial differential equations (see [32]). We put $N_n = \infty$ in the definition 4.1. Let us measure the error of the result in the norm $\eta^{(n)} = 1/m \sum_{j=1}^{m} |\varepsilon_j^{(n)}|$ where $\varepsilon_j^{(n)}$ is the error in one point of the net; m is the number of the net-points and $h = C/n$ is the step. Let the matrix $\boldsymbol{A}$ of finite-difference equations have the form

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}, & \boldsymbol{B} \\ \boldsymbol{B}^T, & \boldsymbol{I} \end{pmatrix}$$

where $\boldsymbol{I}$ is the unit matrix. Then the following theorem holds.

Theorem 4.3. *Let the previous assumptions hold. Then the numerical process is a B_2 process if* $0 < \alpha \leqq \omega \leqq 2 - Dh,\ D > 0$.

Evidently a special case of theorem 4.3 is when ω is independent of h or ω is the optimal overrelaxation parameter.

We shall introduce an example.

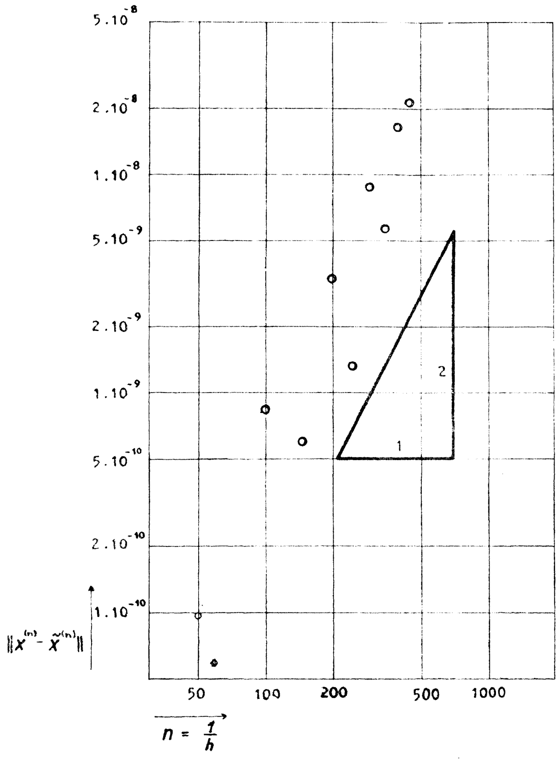


Fig. 4.6

Example 4.2. Let us solve the one dimensional problem $y'' = 1$, $y(0) = y(1) = 0$ with the finite-difference method and overrelaxation. Because of the round-off error the iterations do not, in general, converge to the required solution. They will "quasi-converge" in a more or less well known sense. In Fig. 4.6 we see $\eta^{(n)}$, $n = 1/h$ in dependence on h.

We see a good agreement with theorem 4.3. It is possible to formulate the theorem 4.3 for a $0 < \omega \leqq 1$ in a more general form. See [12].

Further processes have also been investigated. I shall mention here the stability of the Kellog process for the determination of eigenvalues (see [27]) and a numerical process for solving a problems of the theory of reactors [44] and the process for computation of conform mapping. (See also [12.])

I have shown a few different aspect of incredulity with regard to the given information which appear in computations. I think that this kind of investigations is very important when choosing an algorithm in general.

References

[1] *А. А. Абрамов:* О переносе граничных условий для систем линейных обыкновенных диффэренциальных уравнений. Ж. выч. мат. и мат. физ. *1*, 1961, 542—545.

[2] *А. А. Абрамов:* Вариант метода прогонки. Ж. выч. мат. и мат. физ. *1*, 1961, 349—351.

[3] *A. A. Abramov:* Transfer of boundary conditions for system of ordinary linear differential equations. Proc. of IFIP Congress 65, p. 420.

[4] *С. А. Агаханов:* О точности некоторых квадратурных и кубатурных формул. Сиб. мат. журнал *11*, 1965, 1—15.

[5] *I. Babuška:* Über die optimale Berechnung der Fourierischen Koeffizienten. Apl. Mat. *11*, 1966, 113—122.

[6] *I. Babuška:* Über universelloptimale Quadraturformeln. Apl. Mat. 1968.

[7] *I. Babuška, M. Práger:* Numerisch stabile Methoden zur Lösung von Randwertaufgaben. ZAMM *41*, 1961, H. 4—6.

[8] *И. Бабушка, С. Л. Соболев:* Оптимизация численных методов. Apl. mat. *10*, 1965, 96—129.

[9] *I. Babuška, M. Práger, E. Vitásek:* Numerische Stabilität von Rechenprozessen. Wiss. Z. Techn. Hochsch. Dresden *12*, 1963, 101—110.

[10] *I. Babuška, M. Práger, E. Vitásek:* Numerické řešení diferenciálních rovnic 1964, SNTL.

[11] *I. Babuška, M. Práger, E. Vitásek:* Stability of Numerical Processes, Proc. of IFIP 65, 602—603.

[12] *I. Babuška, M. Práger, E. Vitásek:* Numerical Process in Differential Equations. Interscience Publishers 1966.

[13] *F. L. Bauer:* Numerische Abschätzung und Berechnung von Eigenwerten nichtsymmetrischen Matrizen. Apl. Mat. *10*, 1965, 178—189.

[14] *F. L. Bauer et al.* Moderne Rechenanlagen, Stuttgart 1965, p. 64.

[15] *F. L. Bauer:* Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme. ZAMM *46*, 1966, 409—421.

[16] *F. L. Bauer, H. Rutishauser, E. Stiefel:* New Aspect in Numerical Quadrature. Proc. of Symp. in Appl. Mat. 1963, *XV*, 199—218.

[17] *R. Bauman:* Algol Manual der Alcor-Gruppe, Sonderdruck aus Elektronischen Rechenanlagen H 5/6 (1961), H 2 (1962) R. Oldenburg München.

[18] *P. J. Davis:* On the Numerical Integration of Periodic Analytic Functions. Proceedings of Symposium Madison 1959.

[19] *G. G. Dahlquist:* On Rigorous Error Bounds in the Numerical Solution of Ordinary Differential Equations. Numerical Solution of Nonlinear Differential Equations. Wiley 1966, 89—96.

[20] *G. G. Dahlquist:* Private communication.

[21] *H. Ehlich:* Untersuchungen zur numerischen Fourieranalyse. Math. Zeitschr. *91*, 1966, 380—420.

[22] *G. Hämmerlin:* Über ableitungsfreie Schranken für Quadraturfehler. Numerische Mathematik *5*, 1963, 226—233; *7*, 1965, 232—237.

[23] *P. Henrici:* Elements of numerical Analysis. Wiley 1964.

[24] *D. Jagerman:* Investigation of Modified Mid-Point Quadrature Formula, Math. of Comp. *20*, 1966, 78—89.

[25] *G. Kowallewski:* Interpolation und genährte Quadratur. Leipzig 1930, 130.

[26] *В. И. Крылов:* Приближенное вычисление интегралов. Москва 1966.

[27] *I. Marek:* Numerische Stabilität der Prozesse vom Keloggschen Typus. Liblice 1967.

[28] *J. Milota:* Universal Almost Optimal Formulae Solutions of Boundary Value Problems for Ordinary Differential Equations. Liblice 1967. Apl. Mat. *13*, 1968.

[29] *R. E. Moore:* The automatic Analysis and Control of Error in Digital Computation. Vol. *1*, 61—130, Proceedings of a seminar University of Wisconsin, Madison Octobre 5—7, 1964.

[30] *R. E. Moore:* Interval Analysis. Prentice Hall 1966.

[31] *R. E. Moore:* Practical Aspect of Interval Computation. Liblice 1967. Apl. Mat. *13*, 1968, 52—92.

[32] *M. Práger:* Numerical Stability of the Method of Overrelaxation. Liblice 1967.

[33] *P. Přikryl:* On Computation of Fourier Coefficients in Strongly Periodic Spaces. Liblice 1967.

[34] *A. Sard:* Linear Approximation. Providence 1963.

[35] *K. Segeth:* On Universally Optimal Quadrature Formulae Involving Values of Derivatives of Integrand. Liblice 1967.

[36] *С. Л. Соболев:* Лекции по теории кубатурных формул. Новосибирск 1965.

[37] *H. J. Stetter:* Numerical Approximation of Fourier-Transform. Num. Math. 8, 1966, 235—249.

[38] *J. Taufer:* On Factorization Method. Apl. Mat. *11*, 1966, 427—452.

[39] *J. Taufer:* Faktorisierungsmethode für ein Randwertproblem eines linearen Systems von Differentialgleichungen, Liblice 1967, Apl. Mat. *13*, 1968.

[40] *J. Taufer:* Faktorisierungsmethode für ein Eigenwertproblem eines linearen Systems von Differentialgleichungen, Liblice 1967. Apl. Mat. 13, 1968.

[41] *E. Vitásek:* Numerical Stability in Solution of ordinary Differential Equations of Higher Order, Liblice 1967. Apl. Mat. *13*, 1968.

[42] *J. H. Wilkinson:* Rounding errors in algebraic processes. London H.M.S.O. 1963.

[43] *J. H. Wilkinson:* A Survey of Errors Analysis of Matrix Algorithms. Liblice 1967. Apl. Mat. *13*, 1968, 93—102.

[44] *R. Zezula:* Numerische Stabilität eines Algorithmus zur Berechnung des Eigenparameters eines Matrizenoperator mit Hilfe der Reduktionsmethode und der Banachschen Iterationen. Liblice 1967. Apl. Mat. *13*, 1968.

[45] *R. Zurmühl:* Matrizen und ihre technische Anwendungen. Berlin 1964.

Dr. Ing. *Ivo Babuška* DrSc., Matematický ústav ČSAV, Praha 1, Žitná 25, ČSSR.