## EQUADIFF 2

## Ivo Babuška <br> Problems of optimization of numerical mathematics

In: Valter Šeda (ed.): Differential Equations and Their Applications, Proceedings of the Conference held in Bratislava in September 1966. Slovenské pedagogické nakladatel'stvo, Bratislava, 1967. Acta Facultatis Rerum Naturalium Universitatis Comenianae. Mathematica, XVII. pp. 15--25.

Persistent URL: http://dml.cz/dmlcz/700219

## Terms of use:

(C) Comenius University in Bratislava, 1967

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these Terms of use.


This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project DML-CZ: The Czech Digital Mathematics Library http://project.dml.cz

# PROBLEMS OF OPTIMIIZATION OF NUNERICAL MATHEMATICS. 

## I. Babư̌ka, Praha

1. Modern computational techniques are putting forward new problems in numerical analysis. At present numerical mathematics can be considered as a set od constructive mathematical methods transforming given information into desired ones (see e.g. Babuška [1966], Henrici [1964], Babuška, Sobolev [1965], Babuška, Práger, Vitásek [1966]). The classic concepts as for example that of method are begimning to have new meaning. The first place is being occupied by algorithms and the methods are rather comprehend as a class of algorithms of certain kind. Concerning algorithms the following requirements arise:

## a) sufficient generality of algorithms

this requires the algorithm to be applicable to a sufficiently wide class of problems. For example the algorithm of integration by Cotes' formulae of highest order is not sufficiently general as it is applicable only to the narrow class of analytic functions.

## b) Sufficient universal efficiency;

this means that the algorithm should treat the given informations ,,approximately" as well as the optimal algorithm (see below).

## c) suficiently good realizability

By realizability we mean, that the fact, that the computer does not work in the field of real numbers (the rounding off) should not have a great effect
on the result. Especially this is the problem of numerical stability (see Babuška, Práger, Vitásek [1966]).

In this paper we will study some aspects concerning the universal efficiency. In order to illustrate this problem we will restrict us here only to very special cases.
2. Let a Banach space $B$ be given and let $\varphi \in B^{*}$. Our task will be to calculate the value $\varphi(f)$ for a given $f \in B$. The principal idea of (linear) numerical methods of calculation of the value of the functional $\varphi$ is the following. A matrix of functionals $\Phi \equiv\left\{\varphi_{j}^{(n)}\right\}, j=1, \ldots, n, n=1,2, \ldots, \varphi_{j}^{(n)} \in B^{*}$, is given (these functionals will be called calculable functionals). Now it is necessary to construct the functionals $\varphi_{n}=\sum_{j=1}^{n} C_{j}^{(n)} \varphi_{j}^{(n)}$ in such a way that $\varphi_{n}(f) \rightarrow \varphi(f)$ for $n \rightarrow \infty$. In practical cases we take $\varphi_{n}(f) \approx \varphi(f)$ for sufficiently great $n$. There is a number of problems connected with this task.

1) Problem of the estimate for the upper bound of error.

Here the upper bound of the quantity $\varepsilon_{n}\left(\varphi, \varphi_{n} ; B\right)=\left\|\varphi-\varphi_{n}\right\|_{B}{ }^{*}$ is to be estimated.

This problem bears in fact a classic character and is intensively investigated at present (especially it concerns not only the estimate of order, but also of the corresponding constants); in the case of integration of periodic functions see e.g. Sobolev [1965], [1967], Jagerman [1966], Agahanov [1965], Ehlich [1966], Babuška [1965], Ćarušnikov [1966] and others.
2) Problem of the estimate for the lower bound of error.

Here the lower bound of the quantity

$$
\eta_{n}(\varphi, \Phi, B)=\inf _{\alpha_{k}^{(n)}, k=1, \ldots, n}\left\|\varphi-\sum_{k=1}^{n} \alpha_{k}^{(n)} \varphi_{k}^{(n)}\right\|_{B} *
$$

is to be estimated. Also this question is intensively studied at present. See e.g. Sobolev [1965], [1967], Babuška, Sobolev [1965], Bachvalov [1963] and many others. The quantity gives the maximal accuracy at obtainable on the ground of given information.

## 3) Problem of the optimal formula.

The task is to construct the functionals $\varphi_{n}$ in such a way that

$$
\varepsilon_{n}\left(\varphi, \varphi_{n}, B\right)=\eta_{n}(\varphi, \Phi, B)
$$

See e.g. Babuška, Sobolev [1965], Sobolev [1965], [1967], Golomb, Weinberger [1959] etc. The concrete construction of optimal formulae is very difficult and is known only in special cases. In connection with these difficulties formulae are studied, which are asymptotically optimal or optimal by order. See e.g. Babuška, Sobolev [1965], Sobolev [1965]. From the point of view of numerical practice the problem of optimal formulae encounters some difficulties. Beyond the difficulties connected with the construction of optimal formulas there is also the problem of how to choose the space $B$ in a concrete case. We will now illustrate the practical importance of this problem by a simple example.

Let

$$
\varphi(f)=\int_{0}^{1} f(x) \mathrm{d} x
$$

Let $\Phi$ be a matrix of the functional,such that $\varphi_{n+1}(f)=\frac{1}{n} \sum_{s=0}^{n} a_{s}^{(n)} f\left(\frac{s}{n}\right)$ holds. If $\|f\|_{B}^{2}=f^{2}(0)+\int_{0}^{1}\left(f^{\prime}\right)^{2} \mathrm{~d} x$, then the optimal formula will be the trapzzoid-rule. At the same time it is known, that the trapezoid-rule is scarcely used in practice.

The question of how to lower the risk of choosing the space $B$ in a concrete case is the question of universality of the formula.

## 4) Problem of universal optimality by order.

Let $\mathfrak{A}$ be a given system of Banach spaces $B$ embedded in a Banach space $B_{0}$. Let us have a matrix of calculable functionals $\varphi_{j}^{(n)} \in B_{0}^{*}$ and a matrix of coefficients $\Psi=\left\{C_{\left(C_{j}^{(n)}\right.}\right\}, j=1, \ldots, n ; n=1,2, \ldots$. We will use the following notation:

$$
\mathfrak{H}_{\psi^{\Phi}, \varphi}=E\left[B \in \mathfrak{R}, \frac{\left\|\varphi-\sum_{j=1}^{n} C_{j}^{(n)} \varphi_{j}^{(n)}\right\|_{B^{*}}}{\eta_{n}(\varphi, \Phi, B)} \leq C(B)\right]
$$

[where $C(B)$ depends on $B, \varphi, \Phi, \Psi$ but not on $n$ ]. We will say that the formula $\varphi_{n}=\sum_{j=1}^{n} C_{j}^{(n)} \varphi_{j}^{(n)}$ is universally optimal by order with respect to $\mathfrak{A}_{\dot{\psi}}^{\Phi_{i} \varphi . \text { Further }}$
le't us have two formulae given by the matrices $\Psi_{i}=\left\{{ }^{i} C_{j}^{(n)}\right\}, i=1,2, \ldots$ [i.e. $\varphi_{n, i}=\sum_{j=1}^{n}{ }^{i} C_{j}^{(n)} \varphi_{j}^{(n)}$ ]. We will say that the formula given by with the matrix $\Psi_{1}$ is comparable or better or not worse respect to $\mathfrak{V I}$ than the formula
 respectively. The problem of universal optimality lies in
a) characterization of $\mathfrak{H}_{\psi \psi}^{\psi, \varphi}$ for a given formula,
b) characterization of $\mathfrak{H I}$ in such a that the best formula exist,
c) construction of an algorithm leading to this best formula and an estimate of the quantities $\eta_{n}$ and $C(B)$ as functions of $B$.
3. In this part we will give some illustrative assertions concerning the universal optimality. Let us have the task to calculate a functional over the Hilbert space of periodic functions and let us ask, what is (in the intuitive sense) understood the concept of this space. Its intuitive meaning can be perhaps expressed in the following manner.

Definition 1. We will say that a Hilbert space $H$ of $2 \pi$-periodic complex functions has the property $P$, if the following properties are fullfiled.
$P_{1}: H$ is dense in $C_{2: i}$.
$P_{2}$ : if $f \in H$ then also $g(x)=f(x+c) \in H$ for every real $c$ and $\|f\|=\|g\|$. $P_{3}$ : IIs imbedded in $C_{2 i i}$.
Now the following theorem holds.

## Theorem 1. ${ }^{1)}$

Let $H$ have the property 1 . Then

1) $e^{i k x} \in H, \quad k=\ldots,-1,0,1, \ldots$;
2) $\left(e^{i k x}, e^{i l x}\right)=\lambda_{k}^{2} \quad$ for $k=l$ $=0 \quad$ for $k \neq l ;$
3) $\sum_{m=-\infty}^{+\infty} \lambda_{k}^{-2}<\infty$.

It is easy to prove also the inverse theorem.

## Theorem 2.

Let $K$ be the set of all sequences $\Lambda, \Lambda \equiv\left\{\ldots, \lambda_{-1}, \lambda_{0}, \lambda_{1}, \ldots\right\}$ for which $\lambda_{k}>0, k=\ldots,-1,0,1, \ldots$ and $\Sigma \lambda_{k}^{-2}<\infty$.

Let $M$ be a linear space of all trigonometric polynomials with the scalar product $\left(e^{i k x}, e^{i l x}\right)=\lambda_{k}^{2}$ for $k=l$

$$
=0 \quad \text { for } k \neq l
$$

Let $H_{1}$ denote the complete envelope of $M$ in the given norm; then $H_{1}$ has the property $P$.
${ }^{1)}$ It was M. Práger who has drawn my attention to this theorem.

Now we will introduce various systems of spaces with the property $P$. Let $\mathfrak{N l}$ be the system of all Hilbert spaces with the property $P$.
Let $\mathfrak{N}_{1}$ be of all $H_{1}, \lambda \in K_{1} \subset K$, such that if $\Lambda \in K_{1}$, then

1) $\lambda_{k}=\lambda_{-k}, \quad k=0,1,2, \ldots$
2) $\lambda_{k+1} \geq \lambda_{k} \quad k \geq 0$
3) $\sum_{t=0}^{\infty} \dot{\lambda}^{2}[\alpha \bar{n}] \frac{\lambda^{2}[\alpha n]}{+t(2 n+1)} \leq D, \quad 0 \leq x \leq 4$.
l) does not depend on $n$ (but depends on .1.)

Let $\mathfrak{Y N}_{2}$ be the system of all $H_{1}, \Lambda \in K_{2} \subset K_{1}$ satisfying.

$$
\lambda_{k} \leq C+|k|^{1}, \quad \beta>0 .
$$

Now let $\boldsymbol{\Phi} \equiv\left\{\varphi_{j}^{(n)}\right\}, \quad j=1,2, \ldots, 2 n+1, \quad n=1,2, \ldots$

$$
\varphi^{(n)}(f)=f\binom{2 \pi}{2 n+\bar{l}}
$$

be a matrix of calculable functionals and let us turn to the problem of computation of the functional

$$
\varphi(f)=\frac{1}{2 x} \int_{11}^{2 \cdot r} f(x) \dot{\iota}(x) \mathrm{d} \cdot x, \zeta(x) \in L_{2} .
$$

Then the formula becomes ${ }^{2}$ )

$$
q_{n}(f)=\sum_{j=1}^{2 n+1} C\left(\begin{array}{l}
(i) \\
i \\
(i)
\end{array} \psi_{j}^{(i)}\right.
$$

Now the question is how to choose the coefficients $C_{j}^{(n)}\left(\wp_{n}^{\circ}\right)$. The following theorem holds.

## Theorem 3.

A necessary and sufficient rondition that there should exist such $C_{j}^{(n)}(\zeta)$ that the formula

$$
\psi_{n}(f)=\sum_{j=1}^{2 n+1} C_{i}^{(\eta)}(\zeta) \varphi_{j}^{(n)}
$$

should be universally optimal by order with respect to $\mathfrak{N l}$, is that $\zeta(x)$ should be a trigonomotric polynomial. The cosfficients are uniquelly determined except for a finite number of indicess $n$ and are given $b y$

$$
* C_{j}^{(n)}=\frac{1}{2 n+1} \dot{\zeta}\left(\frac{2 \pi}{2 n+1}-j\right)
$$

If $\zeta(x)$ is a more general function, then it follows from theorem 3 that

[^0]a formula, which would be universally optimal by order with respect to $\mathfrak{A}$ does not exist. In connection with what has been said above the question arises whether it is possible to restrict the system of spaces $2 l$ in such a way that universally optimal - by - order formula should exist. This is solved by the following theorem.

Theorem 4.
If $\zeta(x) \in L_{2}$, then $* C_{j}^{(n)}(\zeta)$ exist so that the formula

$$
\varphi_{n}(f)=\sum_{j=1}^{2 n+1} * C_{j}^{(n)}(\zeta) \varphi_{j}^{(n)}
$$

is universally optimal by order with respect to $\mathfrak{9 1}_{1}$. Except for a finite number of indices $n$, the coefficients are uniquelly determined and we have
where

$$
\begin{gathered}
* C_{j}^{(n)}=\frac{1}{2 n+1} S_{n}\left(\frac{2 \pi}{2 n+1} j\right) \\
S_{n}=\sum_{k=-n}^{+n} d_{k} e^{i k x} \quad \text { and } \quad \zeta(x)=\sum_{k=-\infty}^{+\infty} d_{k} e^{i k x}
\end{gathered}
$$

By theorem 3 and 4 the universally optimal - by - order formula is uniquelly determined. It is clear that should we further restrict the system of spaces $\mathfrak{Q}$, then the formula can be determined non uniquelly. In this connection the following theorem holds.

## Theorem 5.

Let $\zeta(x) \in L_{2}$. Then the formula given by theorem 4 is not the only formula universally optimal by order with respect to $\mathfrak{N}_{2}$.

Returning once more to the formula given by theorem 4 we see that it is not optimal in any $H \in \mathfrak{N r}_{1}$ but is universally optimal by order. It is also easy to see that in fact this formula is obtainable by means of the classic (interpolation) method using trigonometric polynomials. From this point of view the connection between the classic (interpolation) theory of quadrature formulae and the theory based on optimization of formulae is well visible. But we will not go further in the study of this problem.

Using the simplest examples, I have given some typical theorems concerning the form of the universal optimality by order. This problem can of course be substantionally extended to include the problem of calculation of functionals as well as operators.
4. In the conclusion let us give some numerical results. Let us compute
for different values of $\alpha$.

$$
I=\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} e^{\alpha \sin x} \cos x \mathrm{~d} x
$$

As the integrand is obviously a $2 \pi$-periodic function, $I$ can be written in the form

$$
I=\int_{-\pi}^{+\pi} e^{\alpha \sin x} \zeta(x) \mathrm{d} x
$$

where $\zeta(x)=\cos x \quad$ for $|x|<\frac{\pi}{2}$

$$
\zeta(x)=0 \quad \text { for } \frac{\pi}{2} \leq x \leq \pi, \quad-\pi \leq x \leq-\frac{\pi}{2}
$$

(here we make use of the symetry of $f=e^{\alpha \sin x}$ with respect to the point $\left.x= \pm \frac{\pi}{2}\right)$. Now the integrand has the form studied in theorem 4. In the following Table together with various formulae the quadrature error is given in dependence on the number of vaules of the function $e^{\alpha \sin x}$ (for $\alpha=1,5,7$ ) used in the calculation. Besides the trapezoid-rule and the Simpson formula also the Romberg formula (see Bauer, Rutishauser, Stiefl [1963]) according to Bauman algorithm [1961] is given under the notation Romberg. Two other modified methods are given as Romberg 1 and Romberg 2. The formula Romberg 1 is that of Bulirsch-Romberg (see Bulirscy [1964]) and the formula Romberg 2 is that of Bulirsch-Stoer (see Bulirsch, Stoer [1965]). The last one is given for comparison although it is not a linearone.

The computation has been carried out on the ICT 1900 with a double precission of word.

## BIBLIOGRAPHY

[1] С. А. АГАХАН О В (1965): Оточности некоторых квадратурных и кубатурных формул. Сиб. мат. ж. VI, 1, 1-15.
[2] I. Babuška (1965): Über optimale Formeln zur numerischen Berechnung linearer Funktionale. Apl. Mat. 441-445.
[3] I. BABUŠKA (1966): О некоторых вопросах оптимизации численных методов матем. анализа. Сб. Некоторые вопросы вычислительной и прикладной математики. Новосибирск 23-39.
[4] I. Babuška, M. Práger, E. Vitásek (1966): Numerical Processes in Differential Equations. Wiley 1966.
[5] I. Babuška, S. L. Sobolev (1965): Оптимизация численных методов. Apl. Mat. 96-13C.
[6] Н. С. БАХ ВАЛ О В, (1936): Об оценке количества өычислительной работы необходимой при приближенном решєнии задач. Доп. к книге Годунов, Рабенький, Введение в теорию разностных схем. Москва.
[7] F. L. Bauer, H. Rutishauser, E. Stiefel (1963): New Aspect in Numerical Quadrature. Proc. Symp. in Appl. Math. 15, Am. Math. Soc.
[8] R. Baumann (1961): Algol-Manual der Alcor Gruppe, Sonderilruck aus Elektronisrhe Rechenanalogen H/ 5/6 (1961) und H 2 (1962). R. Odenbourg, München.
[9] R. Bulinsch (1964): Bemerkung zur Romberg-Integration. Num. Math. No 1.
[10] R. Bulirsch, I. Stoer (1965): Fellerabschätzung und Extrapolation mit rationalen Funktione" bei I'erfahren von Richarlson Typus. Berecht 640.3. München, Math. F.
[11] В. Д. ЧАРУШНИКО В (1966): Об одıой минимаксной задаче, теории кубатурьных формул. ДАН СССР т. 160, ко $1,36-39$.
[12] H. Еhlici (1966): ('ntersuchungen zur numerischen Fourieranalyse. Math. Zeitsch. 9, $380-42 \%$.
[13] M. Coolomb, H. F. Weinberger (1959): Optimal Approximation and Errors Baunas Proceeding of a Symposiam Conducted by the Math. Res. Anter US Army at the Unic. Wisconsin, Marlison Apr. 21-23, 1958. The University of Wisconsin Press.
[14] P. Henrici (1964): Elements of mumerical Analysis. J. Wiley.
[15] D. Jagermann (1966): Investigation of MTorlified. Mid. IPoint Qualrature Formula, Math. of Comp. I'ol. :20, 79-84.
[16] ${ }^{\text {С }}$. Л. СО БОлЕВ: (1965): Лекции по теории кубатурных формул. Новосибирск.
[17] I. BABUŠKA (1968): Problems of optimization and numerical stablíty in computations. Apl. Mat. 388-40.4.
[18] I. BABUS゙にA (1968a): Über universal optimale Quadraturformeln, 304-3.38, 388-$-404$.
[19] K. SEGETH (1968): On universally optimal huadrature formulae involving values oj derivatives of integrand. l'reprint. Math. Inst. df Czechosloval: akademy of Sc.


Table 1. The calculation of I for $\alpha=1$ according to rarious formulae. 1) trapezoid-rule, 2) Simpson formula, 3) universal formula, 4) Romberg formula, 5) Romberg formula 1,
6) Romberg formula 2.

|  | 1 | 2 | 3 | 4 | 5 | 6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | 0.981 | 0.371 | 0.20 1 | -0.95 0 | 0.881 | -0.21 2 |
| 6 | 0.371 | -0.12 1 | 0.110 |  | -0.65 1 | -0.21 2 |
| 8 | 0.201 | -0.61 0 | $0.51-2$ | 0.35-1 | -0.39 0 | 0.151 |
| 10 | 0.131 | -0.21 0 | 0.19 -3 |  |  |  |
| 12 | 0.86 | -0.84-1 | $0.57-5$ |  | 0.390 | 0.240 |
| 14 | 0.630 | -0.41-1 | $0.13-6$ |  |  |  |
| 16 | 0.48 0 | -0.23-1 | 0.25-8 | -0.26-3 | -0.35-1 | -0.48-1 |
| 18 | 0.38 0 | -0.14-1 | 0.39-10. |  |  |  |
| 20 | 0.310 | -0.88-2 | 0.51-12 |  |  |  |
| 22 | 0.250 | -0.59-2 | 0.55-14 |  |  |  |
| 24 | 0.210 | -0.41-2 | 0.49-16 |  | 0.52-3 | -0.56-3 |
| 26 | 0.18 0 | -0.30-2 | -0.26-17 |  |  |  |
| 28 | 0.160 | -0.22-2 | -0.30-17 |  |  |  |
| 30 | 0.140 | -0.17-2 | -0.26-17 |  |  |  |
| 32 | 0.120 | -0.13-2 | 0.00 0 | 0.22-6 | 0.50-4 | 0.44-4 |
| 34 | 0.110 | -0.99-3 | -0.13-17 |  |  |  |
| 36 | 0.94-1 | -0.79-3 | -0.43-18 |  |  |  |
| 38 | 0.85-1 | -0.63-3 | 0.87-18 |  |  |  |
| 40 | 0.76-1 | -0.51-3 | 0.87-18 |  |  |  |
| 42 | 0.69-1 | -0.42-3 | 0.87-18 |  |  |  |
| 44 | 0.63-1 | -0.35-3 | 0.22-17 |  |  |  |
| 46 | 0.58-1 | -0.29-3 | 0.87-18 |  |  |  |
| 48 | 0.53-1 | -0.25-3 | 0.17-17 |  | -0.20-5 | 0.65-8 |
| 50 | 0.49-1 | -0.21-3 | 0.22-17 |  |  |  |
| 52 | 0.45-1 | -0.18-3 | 0.13-17 |  |  |  |
| 54 | 0.42-1 | -0.15-3 | 0.22-17 |  |  |  |
| 56 | 0.39-1 | -0.13-3 | 0.26-17 |  |  |  |
| 58 | 0.36-1 | -0.12-3 | 0.26-17 |  |  |  |
| 60 | 0.34-1 | -0.10-3 | 0.30-17 |  |  |  |
| 62 | 0.32-1 | -0.88-4 | 0.87-18 |  |  |  |
| 64 | 0.30-1 | -0.77-4 | 0.35-17 | 0.45-9 | 0.23-7 | -0.64-9 |

Table 2. The calculation of $I$ for $\alpha=5$ according to various formulae. 1) trapezoid-rule, 2) Simpson formula, 3) universal formula, 4) Romberg formula, 5) Romberg formula 1, 6) Romberg formula 2.


Table 3. The calculation of $I$ for $\alpha=7$ according to various formulae. 1) trapezoid-rule, 2) Simpson formula, 3) universal formula, 4) Romberg formula, 5) Romberg formula 1, 6) Romberg formula 2.


[^0]:    ${ }^{2)}$ To simplify formally the following assertions we have restricted us to an odd number of points used.

