BULLETIN DE L'ACADÉMIE
POLONAISE DES SCIENCES
Cl. III — Vol. IV, No. 5, 1956

MATHEMATICS

Calculus of Approximations

by

M. WARMUS

Presented by H. STEINHAUS on February 6, 1956

This paper presents a theory which lays down the foundations for numerical computations and makes it possible to formulate properly many numerical problems.

By the approximate number $[a, A]$ we shall indicate the interval $[a, A]$, i. e. the set of all real numbers x that satisfy the inequality $a \leqslant x \leqslant A$. The approximate number $[B-b, B+b]$ can also be denoted by $\overset{b}{B}$. Thus, the approximate number $[a, A]$ can be expressed in the form $\frac{1}{2}\overset{\frac{1}{2}(A-a)}{(A+a)}$. We shall omit initial zeros in the upper part, if they lie to the left of the last digit of the lower one. For example, we shall write 3.1416^{02} instead of $3.1416_{0.00002}$.

We say that the approximate numbers α and β are equal and we write $\alpha = \beta$ if, and only if, they are two identical intervals. Hence, we have $[a, A] = [b, B]$ if, and only if, $a = b$ and $A = B$, and similarly $\overset{a}{A} = \overset{b}{B}$ if, and only if, $A = B$ and $a = b$.

We say that the approximate number β approximates the approximate number α and we write $\alpha \Rightarrow \beta$ or $\beta \Leftarrow \alpha$, if, and only if, the interval β includes the interval α. Thus, we have $[a, A] \Rightarrow [b, B]$ if, and only if, $a \geqslant b$ and $A \leqslant B$, and similarly $\overset{a}{A} \Rightarrow \overset{b}{B}$ if, and only if, $b - a \geqslant |B - A|$. It is easy to prove that the approximations $\alpha \Rightarrow \beta$ and $\beta \Rightarrow \gamma$ imply $\alpha \Rightarrow \gamma$.

The relation $\alpha \Rightarrow \beta$ is a partial ordering of the set of all approximate numbers.

In practical computations it is convenient to use the following two rules:

the rounding-off rule: $\overset{a}{A} + c \Rightarrow \overset{a+|c|}{A}$;

the extending rule: $\overset{a}{A} \Rightarrow \overset{b}{A}$ if $b \geqslant a$.

For example,

$$2.7182 \overset{09}{\Rightarrow} 2.72 \overset{0189}{\Rightarrow} 2.72 \overset{02}{.}$$

Here the lower part 2.7182 has been rounded off at first to 2.72 by means of the rounding-off rule, and afterwards the upper part 0.00189 has been rounded off to 0.002 by means of the extending rule.

We say that the approximate number γ is the sum, the difference, the product, or the quotient of the approximate numbers α and β, and we write $\alpha+\beta=\gamma$, $\alpha-\beta=\gamma$, $\alpha\cdot\beta=\gamma$, or $\alpha:\beta=\gamma$ respectively if, and only if, γ is the set of all numbers $x+y$, $x-y$, xy or x/y respectively, where x is a real number from the interval α and y a real number from the interval β. In place of $\alpha:\beta$ we also write $\frac{\alpha}{\beta}$. We assume that the interval β does not include zero, whenever it is a divisor. In that case, if $\beta=[b,B]$, then $bB>0$, and if $\beta=\overset{b}{B}$, then $b<|B|$.

It can be proved that

$$[a,A]+[b,B]=[a+b,A+B], \qquad \overset{a}{A}+\overset{b}{B}=\overset{a+b}{A+B},$$

$$[a,A]-[b,B]=[a-B,A-b], \qquad \overset{a}{A}-\overset{b}{B}=\overset{a+b}{A-B},$$

$$[a,A]\cdot[b,B]=[\min(ab,aB,Ab,AB),\ \max(ab,aB,Ab,AB)],$$

$$\overset{a}{A}\cdot\overset{b}{B}=\operatorname{sgn}(AB)[|AB|+\min(|A|b,a|B|,ab)]^{|A|b+a|B|+ab-\min(|A|b,a|B|,ab)},$$

$$[a,A]:[b,B]=\left[\min\left(\frac{a}{b},\frac{a}{B},\frac{A}{b},\frac{A}{B}\right),\ \max\left(\frac{a}{b},\frac{a}{B},\frac{A}{b},\frac{A}{B}\right)\right],$$

$$\overset{a}{A}:\overset{b}{B}=\overset{a}{A}\cdot\overset{b}{B}\cdot\frac{\overset{0}{1}}{B^2-b^2}.$$

For example,

$$(-\overset{2}{13})\cdot\overset{7}{3}=(-\overset{91+14}{(39+6)})=(-\overset{105}{45}).$$

Approximate numbers do not form a group with respect to addition, because the equation $\overset{a}{A}+\xi=\overset{b}{B}$ has no solution whenever $b<a$. Moreover, subtraction is not the inverse operation of addition, that is, the equality $(\overset{a}{A}+\overset{b}{B})-\overset{b}{B}=\overset{a}{A}$ does not hold whenever $b\neq 0$, but $(\overset{a}{A}+\overset{b}{B})-\overset{b}{B}=\overset{a+2b}{A}$.

Similarly, division is not the inverse operation of multiplication. It is a consequence of the irreversible process of error accumulation.

Addition and multiplication are both commutative and associative, but the distributive laws fail for multiplication and division. However, the relations

$$\alpha\cdot(\beta+\gamma)\Rightarrow\alpha\cdot\beta+\alpha\cdot\gamma, \qquad \alpha\cdot(\beta-\gamma)\Rightarrow\alpha\cdot\beta-\alpha\cdot\gamma,$$

$$(\alpha+\beta):\gamma\Rightarrow\alpha:\gamma+\beta:\gamma, \qquad (\alpha-\beta):\gamma\Rightarrow\alpha:\gamma-\beta:\gamma$$

hold. Therefore, we say that multiplication and division are weakly distributive with respect to addition and subtraction.

We can also multiply or divide approximate numbers in an approximate number by using the following formulas:

$$\overset{a}{A}\cdot\overset{b}{B}\Rightarrow\overset{|A|b+a|B|+ab}{AB}, \qquad \overset{a}{A}:\overset{b}{B}\Rightarrow\overset{\frac{a+\left|\frac{A}{B}\right|b}{|B|-b}}{\frac{A}{B}};$$

these formulas are very convenient in practical computations.

The set of all approximate numbers, the upper part of which equals zero, that is, the set of approximate numbers which are intervals reduced to points, is isomorphic to the set of all real numbers with respect to the arithmetical operations. We shall denote the isomorphism by writing $\overset{0}{A}=A$.

The definition of approximate numbers and of the operations on them given above are not sufficient in practical computations. Although the operations were defined in a natural way, they are not regular enough (multiplication and division are not distributive with respect to addition and subtraction) and the inverse operations fail. Therefore we shall now extend the conception of an approximate number and introduce some new operations in order to obtain a ring of approximate numbers.

From now on we shall indicate by the approximate number $[a,A]$ the interval $[\min(a,A),\max(a,A)]$ with the sense from a to A or reduced to a point, if $a=A$. Thus, there is now no need to assume $a\leqslant A$. The approximate number $[B-b,B+b]$ may also be denoted by $\overset{b}{B}$, but now $b\geqslant 0$ is not assumed.

If $a<A$, we shall call the approximate number $[a,A]$ a positively oriented interval or simply an interval. If $a\geqslant A$, it will be called a negatively oriented interval. Thus $\overset{b}{B}$ is an interval if, and only if, $b\geqslant 0$. If $b\leqslant 0$, $\overset{b}{B}$ is a negatively oriented interval. We shall identify the appro-

ximate numbers in the previous sense with the intervals in the new sense. All the natural operations will hold for the intervals.

We now define equality, approximation and the four regular arithmetical operations for all approximate numbers as follows:

$$[a,A]=[b,B] \text{ if, and only if, } a=b,\ A=B, \text{ i. e. } \overset{a}{A}=\overset{b}{B} \text{ if, and only if, } A=B,\ a=b,$$

$$[a,A]\Rightarrow[b,B] \text{ if, and only if, } a\geqslant b,\ A\leqslant B, \text{ i. e. } \overset{a}{A}\Rightarrow\overset{b}{B} \text{ if, and only if, } b-a\geqslant|B-A|,$$

$$[a,A]+[b,B]=[a+b,A+B], \quad \text{i. e.} \quad \overset{a}{A}+\overset{b}{B}=\overset{a+b}{A+B},$$

$$[a,A]-[b,B]=[a-b,A-B], \quad \text{i. e.} \quad \overset{a}{A}-\overset{b}{B}=\overset{a-b}{A-B},$$

$$[a,A]\cdot[b,B]=[ab,AB], \quad \text{i. e.} \quad \overset{a}{A}\cdot\overset{b}{B}=\overset{aB+Ab}{AB+ab},$$

$$[a,A]:[b,B]=\left[\frac{a}{b},\frac{A}{B}\right] \text{ if } b\neq0,\ B\neq0, \quad \text{i. e.} \quad \overset{a}{A}:\overset{b}{B}=\overset{\frac{aB-Ab}{B^2-b^2}}{\frac{AB-ab}{B^2-b^2}} \text{ if } |B|\neq b.$$

We see that the natural addition of intervals is identical with regular addition.

Approximate numbers form a ring with respect to addition and regular multiplication. Regular subtraction and division are the inverse operations of addition and regular multiplication respectively. All the approximate numbers $[0,A]$ and $[a,0]$, i. e. all the approximate numbers $\overset{a}{A}$ with $|A|=|a|$, are divisors of zero in the ring, but they do not form an ideal, because every approximate number is a sum of two divisors of zero: $[a,A]=[a,0]+[0,A]$.

Every approximate number is a linear combination of the approximate numbers $\overset{0}{1}$ and $\overset{1}{0}$. Therefore, approximate numbers form a two-dimensional linear vector space with respect to addition, and regular multiplication by real numbers.

The set of all approximate numbers $\overset{0}{A}$ is isomorphic to the set of all real numbers with respect to the regular arithmetical operations. We denote the isomorphism, as before, by writing $\overset{0}{A}=A$. Every approximate number $\overset{a}{A}$ can be written in the form $A+a\Lambda$, where $\Lambda=\overset{1}{0}$ ($\Lambda^2=\Lambda\Lambda=1$). Thus we may compute with approximate numbers (excepting the division by a divisor of zero) in a similar manner as with complex numbers. For example:

$$\overset{3}{5}\cdot\overset{-2}{6}=(5+3\Lambda)(6-2\Lambda)=30-10\Lambda+18\Lambda-6\Lambda^2=\overset{8}{24}+8\Lambda=\overset{8}{24}.$$

Approximate numbers can be represented by points in a plane. Let (X,x) be the rectangular Cartesian co-ordinates of a point P in the plane. We say that the point P represents the approximate number $X+x\Lambda$. In such a way we have a one-to-one correspondence between the approximative numbers and the points on a plane.

If $|a|<|A|$, then the approximate number $\overset{a}{A}$ can be written in the form $\varrho(\operatorname{ch}\psi+\Lambda\operatorname{sh}\psi)$, where the modulus ϱ and the argument ψ are real, $\varrho=\pm\sqrt{A^2-a^2}$. We call this the hyperbolic form of the approximate number $\overset{a}{A}$. If $|a|>|A|$, ϱ and ψ are complex. If $|a|=|A|$, that is if $\overset{a}{A}$ is a divisor of zero, it cannot be written in a hyperbolic form. It is easy to verify that, as with complex numbers in the trigonometric form, the modulus of the regular product of two approximate numbers is the product of their moduli and the argument of the product is the sum of the arguments. The modulus of the regular quotient of two approximate numbers is the quotient of their moduli and the argument of the quotient is the argument of the numerator minus the argument of the denominator. As in De Moivre's theorem, if n is a positive integer, then $[\varrho(\operatorname{ch}\psi+\Lambda\operatorname{sh}\psi)]^n=\varrho^n(\operatorname{ch}n\psi+\Lambda\operatorname{sh}n\psi)$.

All the natural operations can be translated into the regular ones. For example,

$$\overset{a}{A}:\overset{b}{B}=\begin{cases} \overset{a\operatorname{sgn}B}{A} : \overset{-b\operatorname{sgn}A}{B} & \text{if } |A|\geqslant a\geqslant0,\ |B|\geqslant b\geqslant0, \\ \overset{a\operatorname{sgn}B}{A} : \overset{}{(B-b\operatorname{sgn}B)} & \text{if } |A|\leqslant a,\quad |B|\geqslant b\geqslant0. \end{cases}$$

In many numerical problems it is convenient to operate in the following way. Formulate the problem by use of the natural operations; afterwards translate these operations into the regular ones, and then solve the problem by their use.

Many numerical problems reduce to a system of approximations. We compute with approximations in a manner similar to that used for equations or inequalities in the usual algebra.

In order to introduce a topology into the ring of approximate numbers we now define a norm. By the norm of an approximate number $a=\overset{a}{A}$ we mean the real number $\|a\|=|A|+|a|$. Such a norm has all the

required properties and besides $\|\alpha\cdot\beta\|\leqslant\|\alpha\|\cdot\|\beta\|$. Thus, approximate numbers form a two-dimensional Banach algebra with respect to addition, regular multiplication and regular multiplication by real numbers.

A sequence of approximate numbers $\{a_n\}$ is said to be convergent to a if, and only if, it converges in the norm, i. e. $\lim_{n\to\infty}\|a_n-a\|=0$. Thus, it follows that

$$\lim_{n\to\infty}[a_n,A_n]=[\lim_{n\to\infty}a_n,\lim_{n\to\infty}A_n] \quad\text{and}\quad \lim_{n\to\infty}\overset{a_n}{A_n}=\overset{\lim a_n}{\lim_{n\to\infty}A_n}.$$

We shall call a function with real arguments and approximate values, an approximate function. For example, let $F(x)$, $f(x)$ and $G(x)$ be three usual functions of the real variable x, such that $|G(x)-F(x)|\leqslant f(x)$. Then we say that the approximate function $\overset{f(x)}{F}(x)$ approximates the function $G(x)$ and we write

$$G(x)=\Rightarrow\overset{f(x)}{F}(x).$$

We define continuity, convergence, derivative and integral of an approximate function $\varphi(x)=\overset{f(x)}{F}(x)$ in the obvious way, and the following formulas result:

$$\varphi'(x)=\overset{f'(x)}{F'}(x), \qquad \int_a^b\varphi(x)\,dx=\overset{\int_a^b f(x)\,dx}{\int_a^b F(x)\,dx}.$$

If $\overset{f(x)}{F}(x)=\Rightarrow\overset{g(x)}{G}(x)$ and $g(x_0)=0$, then $\overset{f'(x_0)}{F'}(x_0)=\Rightarrow\overset{g'(x_0)}{G'}(x_0)$.

If $\varphi(x)=\Rightarrow\psi(x)$, then $\int_a^b\varphi(x)\,dx=\Rightarrow\int_a^b\psi(x)\,dx$.

We shall call a function with approximate arguments and values a function of approximate variables. For example, the approximate value of the integral $\int_x^X f(u)\,du$, where $f(u)$ is a fixed usual function of a real variable u, is a function of the interval $[x,X]$, if the method of computing is uniquely defined.

Continuity, convergence, derivatives and integrals of functions of approximate variables are introduced in a manner similar to the analogous conceptions for functions of complex variables. They have many

interesting properties, which, however, seem to be of no great use in numerical problems.

The author of this paper has elaborated the details of the theory outlined above. He is preparing a monograph. All the numerical methods and computations can be written in the language of this theory; many new methods arise; many cumbersome computations can be performed automatically. There are many numerical problems which cannot be formulated properly, because this theory fails.

INSTITUTE OF MATHEMATICS, POLISH ACADEMY OF SCIENCES