ON THE EXPANSION OF π IN A REGULAR CONTINUED FRACTION

PEDER PEDERSEN

Let

$$\pi = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \dots}},$$

where $a_0, a_1, a_2, \dots$ are positive integers, be an expansion of π in a regular continued fraction. The unknown integers $a_0, a_1, a_2, \dots$ can be computed from a representation of π by a decimal fraction. In 1685 J. Wallis [1] computed the first 34 partial quotients of π from the 35 first decimals of π. A century later Lambert [2] recomputed the partial quotients of π, and found a sequence of quotients which agreed with the one found by Wallis up to the 26th quotient only. There were, however, indications that the partial quotients found by Wallis were correct [3, p. 42, 62].

During the 19th century, no one attempted to settle whether Wallis' or Lambert's computation was correct. In 1938, however, the expansion of π in a continued fraction was reconsidered by D. H. Lehmer in [4] and [5]. In the first of these papers, Lehmer uses a decimal expansion of π to 100 decimals and computes the partial quotients up to a_{90}. In the second paper, the partial quotients are computed up to a_{100}. A comparison shows that all the partial quotients given by Wallis were correct except a_{34}, which Wallis found to be 1, whereas the correct value is 99.

An interesting feature in the first paper by Lehmer is a method by which the labour involved in the Euclidean algorithm is considerably reduced. The partial quotients are computed in 5 steps, each step including about 20 quotients. The partial quotients in the first step can be determined by using only the first decimals in the expansion of π. From the partial quotients in the first step, the corresponding convergents are computed, and these are then used for determining two numbers, of which the first digits are used in an Euclidean algorithm for the computation of partial quotients in the second step. We proceed similarly from the second step to the third step, and so forth till we have computed as many partial quotients as is possible from the chosen number of decimals in π. The advantage of the method can be characterized as follows: troublesome divisions by large numbers are partly replaced by multiplications, which can be more easily carried out on a calculating machine.

Without knowing Lehmer's work, and prompted by a remark by Perron [3, p. 42] about the discrepancy between the values of the partial quotients found by Wallis and Lambert, the author in 1945 developed π in a regular continued fraction, starting with an approximation of π to 100 decimals. Using the Euclidean algorithm, all partial quotients up to a_{96} were computed; a_{96} being the last quotient which can be determined from the chosen value of π. The result of the computation is given in [6].

Only recently my attention was called to the papers by Lehmer from 1938 and 1939. It can be added that there is complete agreement between the values of the partial quotients found by Lehmer and myself.

The first paper by Lehmer deals exclusively with the computation of the partial quotients in a continued fraction. The main part of the second paper contains a computation of Khintchine's constant. At the end of this second paper, Lehmer gives the partial quotients $a_{91}, a_{92}, \ldots, a_{100}$, so that all the partial quotients in the continued fraction expansion of π up to a_{100} are known. In an interesting manner, Lehmer then uses the computed values for a computation related to Khintchine's and Lévy's constants.

For the study of the constants mentioned above, it is thus of some importance to know the partial quotients in such cases where they behave rather irregularly, and I have therefore in the present paper computed the partial quotients in the continued fraction expansion of π up to a_{200}.

The Euclidean algorithm is the basis for the expansion in continued fraction of a given number. The exposition given here follows in the main the exposition of Lehmer.

Let x_0 and x_1 be two positive numbers. The Euclidean algorithm belonging to the number x_0/x_1 is

$$\begin{aligned} x_0 &= a_0x_1+x_2 \\ x_1 &= a_1x_2+x_3 \\ &\cdots\cdots\cdots\cdots \\ x_n &= a_nx_{n+1}+x_{n+2} \\ &\cdots\cdots\cdots\cdots\cdot \end{aligned} \tag{1}$$

Here a_n is the largest integer not exceeding x_n/x_{n+1}. Eliminating $x_2, x_3, \dots$ from (1), we obtain the regular continued fraction for ξ_0:

$$\xi_0 = \frac{x_0}{x_1} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \dots}} = [a_0, a_1, a_2, \dots] .$$

Denoting the convergents by A_n/B_n, we have

$$\frac{A_n}{B_n} = [a_0, a_1, \dots, a_n] ,$$

where A_n and B_n can be computed from the recursion formulas

$$A_\nu = a_\nu A_{\nu-1} + A_{\nu-2}$$
$$B_\nu = a_\nu B_{\nu-1} + B_{\nu-2} ,$$

starting with the initial values

$$A_{-1} = 1, \quad A_0 = a_0, \quad B_{-1} = 0, \quad B_0 = 1 .$$

By ξ_ν we denote a continued fraction formed from the partial quotients which appear in the continued fraction expansion for ξ_0, having left out the first ν partial quotients. We then have

$$\xi_\nu = \frac{x_\nu}{x_{\nu+1}} = [a_\nu, a_{\nu+1}, a_{\nu+2}, \dots] .$$

The convergents in the continued fraction expansion of ξ_ν will be denoted by $A_{n,\nu}/B_{n,\nu}$, setting

$$\frac{A_{n,\nu}}{B_{n,\nu}} = [a_\nu, a_{\nu+1}, \dots, a_{\nu+n}] .$$

Here $A_{n,\nu}$ and $B_{n,\nu}$ can be computed from the recursion formulas

$$A_{n,\nu} = a_{\nu+n} A_{n-1,\nu} + A_{n-2,\nu}$$
$$B_{n,\nu} = a_{\nu+n} B_{n-1,\nu} + B_{n-2,\nu} ,$$

using

$$A_{-1,\nu} = 1, \quad A_{0,\nu} = a_\nu, \quad B_{-1,\nu} = 0, \quad B_{0,\nu} = 1$$

as initial values.

A formula which is important for the following computations can be obtained by eliminating $x_{\nu+2}, x_{\nu+3}, \dots, x_{\nu+n-1}$ from (1). As a result of the elimination, we obtain

$$x_{\nu+n} = (-1)^n \{B_{n-2,\nu} x_\nu - A_{n-2,\nu} x_{\nu+1}\} . \tag{2}$$

In particular, we obtain for $\nu=0$ a formula which expresses x_n by x_0 and x_1:

$$x_n = (-1)^n\{B_{n-2}x_0 - A_{n-2}x_1\} . \tag{3}$$

We finally mention the two formulas which establish a relation between the convergents in the continued fraction expansions for ξ_0 and ξ_ν:

$$\begin{aligned} A_{n+\nu-1} &= A_{\nu-1}A_{n-1,\nu} + A_{\nu-2}B_{n-1,\nu} \\ B_{n+\nu-1} &= B_{\nu-1}A_{n-1,\nu} + B_{\nu-2}B_{n-1,\nu} . \end{aligned} \tag{4}$$

As a basis for the computation of the partial quotients $a_{97}, a_{98}, \ldots, a_{200}$, we use the denominators B_{95} and B_{96} which were found earlier [6], and a value of π to 220 decimal places. A preliminary estimate shows that this number of decimals will suffice, and this is confirmed by the actual computation.

The first step consists of finding the equation in the Euclidean algorithm (1), from which a_{97} is to be determined. This equation has the form

$$x_{97} = a_{97}x_{98} + x_{99} .$$

From this equation it appears that x_{97} and x_{98} must be calculated in order that a_{97} can be found. But from (3) we get for $n=97$ and 98:

$$\begin{aligned} x_{97} &= -B_{95}x_0 + A_{95}x_1 \\ x_{98} &= B_{96}x_0 - A_{96}x_1 . \end{aligned}$$

Let π_{220} be an approximation to π to 220 decimals. The fraction x_0/x_1 being equal to π_{220}, we can put $x_1=1$, hence $x_0=\pi_{220}$. The expressions for x_{97} and x_{98} then become

$$\begin{aligned} x_{97} &= -B_{95}\pi_{220} + A_{95} \\ x_{98} &= B_{96}\pi_{220} - A_{96} . \end{aligned}$$

The numbers A_{95}, B_{95}, A_{96} and B_{96}, which appear in the above expressions, are of the order of magnitude 10^{50}. On the other hand, x_{97} and x_{98} are of the order of magnitude 10^{-50}. From this fact it follows in the first place that A_{96} is the largest integer less than $B_{96}\pi_{220}$, and secondly that in the number $B_{96}\pi_{220}$ about the first 50 digits after the decimal point are all equal to 0, and after these 0's begin the digits which determine the value of x_{98}.

Something similar holds for the product $B_{95}\pi_{220}$. The largest integer less than $B_{95}\pi_{220}$ is $A_{95}-1$, and in the number $B_{95}\pi_{220}$ about the first 50 digits after the decimal point are all equal to 9, and after these 9's begin the digits which determine the value of x_{97}.

Thus it appears that the values of A_{95} and A_{96} are not needed for the computation of x_{97} and x_{98}.

The difficulty in finding x_{97} and x_{98} lies in computing the two products $B_{95}\pi_{220}$ and $B_{96}\pi_{220}$. Each of the numbers B_{95} and B_{96} contains 50 digits, and π_{220} contains 221 digits. The labour involved in carrying out the multiplication can be shortened considerably, since it is not necessary to compute the digits before the decimal point, nor is it necessary to compute the first 50 digits after the decimal point, since these digits are either 0 or 9. Altogether we can disregard about 100 digits in the product. The calculations can therefore be arranged in such a way that only the significant decimals in x_{97} and x_{98} are found. It may still be of advantage to let the computation overlap some of the 0's or 9's and in this way obtain a valuable check.

As a result of the computation, we find

$$x_{97} = 0.00000\ldots00001\ 58223\ldots$$
$$x_{98} = 0.00000\ldots00000\ 82722\ldots,$$

where x_{97} contains 49 zeros and x_{98} contains 50 zeros after the decimal point. The total number of decimals in each of the two numbers is 170.

The Euclidean algorithm belonging to x_0/x_1 can now be continued by using x_{97} and x_{98} as initial values. But instead of working with all significant digits in x_{97} and x_{98} (121 and 120 digits respectively), we shall, following the method given by Lehmer, proceed in a sequence of small steps. In the first step we work with the integers

$$[x_{97}] = 1\ 58223\ 36281\ 52479\ 34447$$
$$[x_{98}] = 82722\ 03146\ 50186\ 85067\,,$$

that is, we use the first 21 significant digits in x_{97} and the first 20 significant digits in x_{98}.

From the Euclidean algorithm based on these numbers, we can find the 20 partial quotients from a_{97} up to a_{116}. These quotients are

$$1,\ 1,\ 10,\ 2,\ 5,\ 4,\ 1,\ 2,\ 2,\ 8,\ 1,\ 5,\ 2,\ 2,\ 26,\ 1,\ 4,\ 1,\ 1,\ 8\,.$$

From this sequence of partial quotients we determine a sequence of convergents belonging to the continued fraction

$$\xi_{97} = [1,\ 1,\ 10,\ \ldots]\,.$$

Thus we find

$$(5)\qquad \begin{aligned} A_{18,97} &= 7337\ 17748, & B_{18,97} &= 3836\ 00889 \\ A_{19,97} &= 62701\ 74397, & B_{19,97} &= 32781\ 60409\,. \end{aligned}$$

This marks the end of the first step in the computation of the partial quotients.

As an introduction to the next step we must compute x_{117} and x_{118}. From equation (2), with $\nu = 97$ and $n = 20$ and 21, we obtain

$\alpha \backslash \beta$	0	1	2	3	4	5	6	7	8	9
0	3	7	15	1	292	1	1	1	2	1
1	3	1	14	2	1	1	2	2	2	2
2	1	84	2	1	1	15	3	13	1	4
3	2	6	6	99	1	2	2	6	3	5
4	1	1	6	8	1	7	1	2	3	7
5	1	2	1	1	12	1	1	1	3	1
6	1	8	1	1	2	1	6	1	1	5
7	2	2	3	1	2	4	4	16	1	161
8	45	1	22	1	2	2	1	4	1	2
9	24	1	2	1	3	1	2	1	1	10
10	2	5	4	1	2	2	8	1	5	2
11	2	26	1	4	1	1	8	2	42	2
12	1	7	3	3	1	1	7	2	4	9
13	7	2	3	1	57	1	18	1	9	19
14	1	2	18	1	3	7	30	1	1	1
15	3	3	3	1	2	8	1	1	2	1
16	15	1	2	13	1	2	1	4	1	12
17	1	1	3	3	28	1	10	3	2	20
18	1	1	1	1	4	1	1	1	5	3
19	2	1	6	1	4	1	120	2	1	1
20	3									

Table I

$$x_{117} = B_{18,97}x_{97} - A_{18,97}x_{98}$$
$$x_{118} = A_{19,97}x_{98} - B_{19,97}x_{97}\ .$$

The values for x_{117} and x_{118} are now used in a similar way as x_{97} and x_{98}, and it turns out that the partial quotients from a_{117} to a_{133} can be found.

In the third step we then find the partial quotients from a_{134} to a_{148}, in the fourth, fifth and sixth steps we find the partial quotients from a_{149} to a_{171}, from a_{172} to a_{191}, and from a_{192} to a_{200} respectively.

The result of the computation is given in table I, which contains all partial quotients from a_0 to a_{200}. The partial quotient a_n is here denoted by $a_{\alpha\beta}$, where β is the last digit of n and α stands for the number formed by the remaining digits of n.

This solves in a way our problem, but in order to check our computations, it is important to compute at least some of the convergents belonging to the continued fraction.

As initial values we use A_{95}, B_{95}, A_{96} and B_{96}. Using equations (4) and putting $\nu = 97$ and $n = 19$, we obtain

$$A_{115} = A_{96}A_{18,97} + A_{95}B_{18,97}$$
$$B_{115} = B_{96}A_{18,97} + B_{95}B_{18,97}\ ,$$

and by putting $\nu = 97$ and $n = 20$, we obtain

$$A_{116} = A_{96}A_{19,97} + A_{95}B_{19,97}$$
$$B_{116} = B_{96}A_{19,97} + B_{95}B_{19,97}\ .$$

The four numbers $A_{18,97}$, $B_{18,97}$, $A_{19,97}$ and $B_{19,97}$ are precisely those given in (5), and it is therefore possible to compute A_{115}, B_{115}, A_{116} and B_{116} from the four equations given above.

From the four numbers A_{115}, B_{115}, A_{116} and B_{116}, we find in a similar way A_{132}, B_{132}, A_{133} and B_{133}. Continuing this way, we finally obtain A_{199}, B_{199}, A_{200} and B_{200}. The values worked out are given below.

A_{115} = 14568 28022
05181 73875 12246 15843 96739 57004 73722 68189 47377 34606

B_{115} = 4637 22761
88867 08791 99782 01502 88713 00051 50363 42235 41866 16619

A_{116} = 1 24496 99887
45789 04010 63661 55256 01511 49764 54553 33790 60338 90313

B_{116} = 39628 62554
19907 26261 22622 86579 00463 63439 12658 94998 43510 74158

$A_{132} =$ 21775 21346 97428 27726
31896 39360 46693 45743 67244 83827 61244 07906 80231 81267

$B_{132} =$ 6931 26572 11815 85049
36450 38073 47336 85726 20506 56576 77981 45125 96705 39277

$A_{133} =$ 28055 66453 91640 26167
88439 45381 72109 17066 48754 64958 95963 00532 06925 51644

$B_{133} =$ 8930 39538 62719 13130
89158 59654 96062 65611 47285 80844 63970 64369 05358 52361

$A_{147} =$ 33530 04282 58231 73538 78651 88641
58246 51900 33381 68889 97106 64194 82440 64617 94760 34521

$B_{147} =$ 10672 94411 56254 01312 93749 87921
27901 73514 28027 01913 64783 09561 77864 73760 07102 63755

$A_{148} =$ 65983 26041 76767 96185 20427 18624
07583 21000 36380 70125 02108 22221 06439 66505 40149 73630

$B_{148} =$ 21003 12411 35861 21120 97382 00566
32100 56578 25906 35277 65240 51505 17078 94238 40414 50531

$A_{170} =$ 19562 67991 34289 83460 60076 99100 22310 99567
00865 86937 56307 01763 53532 87366 28209 14526 50154 45601

$B_{170} =$ 6226 99441 66935 07672 80722 53361 92506 70294
07684 77540 95676 57488 63083 13774 03311 86060 33116 85792

$A_{171} =$ 37710 43527 51232 28591 46329 22751 86474 13279
78958 40989 43051 33802 64505 07634 43268 42102 33392 18920

$B_{171} =$ 12003 60436 03656 79752 09703 64381 92216 03095
07896 44674 76100 37258 75142 68960 55964 86127 56884 86239

$A_{190} =$ 6940 71435 62613 24291 13895 79783 24190 96687 67041 34606
08890 00522 53605 98011 40594 57443 02273 41726 92661 49443

$B_{190} =$ 2209 29799 67757 44073 81056 51602 66558 67328 74660 99680
35614 39620 25124 10083 91413 62832 99148 16960 43584 77235

$A_{191} =$ 9939 18757 03585 90019 23176 29777 82467 00101 69479 30758
58485 30145 43118 17735 63125 82986 08949 78753 85127 73005

$B_{191} =$ 3163 74166 42801 89726 23843 75688 52136 34740 85441 88952
81979 06676 97836 97039 70662 41968 26293 70217 41022 18418

$A_{199} =$ 2722
52067 61227 41156 04820 16369 10149 52455 80376 29127 74696
09660 74719 37893 33391 45590 90569 88538 51544 04202 44451

$B_{199} =$ 866
60524 65496 46562 72926 67594 90027 66400 54365 43883 92932
73326 35120 20743 05984 73820 18860 52979 28184 01138 27831

$A_{200} =$ 9800
17613 20888 23621 50045 52208 16620 23582 32495 54679 95158
24531 21170 18519 77163 37781 96634 02622 29704 65887 62051

$B_{200} =$ 3119
49294 91862 95341 33892 30555 01530 51653 50046 07765 45528
29009 24137 22262 93718 76624 48358 56139 25327 15414 16482

An extensive calculation like the present one calls for a very thorough check of the final result. The partial quotients and the four numbers A_{199}, B_{199}, A_{200} and B_{200} have therefore been checked in the following way.

Assuming that the A's and B's are computed correctly from the partial quotients, one can obtain a check of the latter by using a theorem which is due to Legendre [3, pp. 44–45]. The fraction A_{200}/B_{200} is a convergent in the continued fraction expansion of π, if and only if it satisfies the inequalities

$$-\frac{1}{B_{200}(2B_{200}-B_{199})} < \pi - \frac{A_{200}}{B_{200}} < \frac{1}{B_{200}(B_{200}+B_{199})}.$$

Using the inequalities in this form is rather inconvenient because it implies the carrying out of the arduous division A_{200}/B_{200}. We therefore rewrite the inequalities by multiplying by B_{200}; we then obtain the inequalities

$$-\frac{1}{2B_{200}-B_{199}} < \pi B_{200} - A_{200} < \frac{1}{B_{200}+B_{199}}.$$

Since

$$-\frac{1}{2B_{200}-B_{199}} = -1.86\cdot 10^{-104},$$

$$\pi B_{200} - A_{200} = 2.43\cdot 10^{-104},$$

and

$$\frac{1}{B_{200}+B_{199}} = 2.51\cdot 10^{-104},$$

we find that the inequalities are satisfied.

The four numbers A_{199}, B_{199}, A_{200} and B_{200} may be checked by the fact that they must satisfy the equation

$$A_{199}B_{200} - A_{200}B_{199} = 1\,.$$

A direct check of this equation is rather arduous, and such a check has not been performed here. One can obtain a rather effective check of the equation by using a remainder test, which requires much less work. In testing, we use the fact that the remainder on dividing 10^{pq} by 10^p-1 is 1. To determine the principal remainder on dividing a number t by 10^p-1, we group the digits of t beginning from the last digit in sections of p digits. Each group is considered as a p-digit number, and the sum of these numbers is a remainder on dividing t by 10^p-1.

In the actual check I used $p=10$. It is then rather easy to determine the principal remainders on dividing A_{199}, B_{199}, A_{200} and B_{200} by $10^{10}-1$. Replacing A_{199}, B_{199}, A_{200} and B_{200} in the expression $A_{199}B_{200}-A_{200}B_{199}$ by their respective principal remainders, we obtain a remainder modulo $10^{10}-1$, and from this we finally find the principal remainder modulo $10^{10}-1$. The check shows that this principal remainder is 1.

Khintchine [7] has proved the following theorem:

The geometric mean $\sqrt[n]{a_1a_2\ldots a_n}$ *of the partial quotients of a regular continued fraction* ξ *converges for almost all* ξ *for* $n\to\infty$ *to the constant*

$$\prod_{r=1}^{\infty}\left(1+\frac{1}{r(r+2)}\right)^{\frac{\log r}{\log 2}}.$$

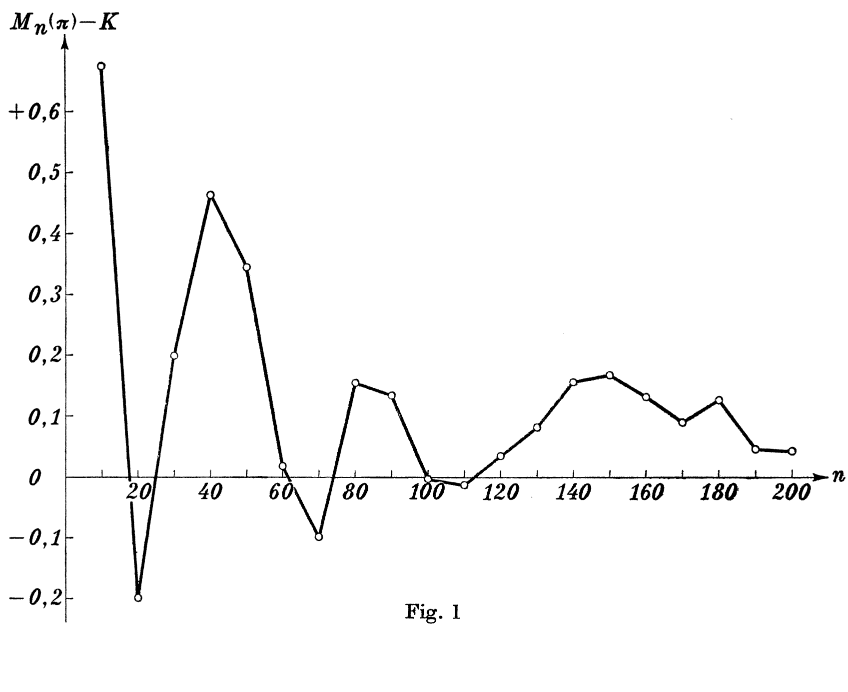

Fig. 1

Khintchine gives the value of the constant as 2.6..., but there are certain difficulties involved in computing the constant because of the rather slow convergence of the infinite product. Lehmer [5], however, has shown how the constant can be computed with great accuracy, and he found the value

$$K = 2.685550 .$$

Having computed this constant K, Lehmer used the opportunity to compare the geometric mean $M_{100}(\pi)$ of the first 100 partial quotients in the continued fraction expansion of π with K. In this way he found that

$$M_{100}(\pi) = 2.683147 ,$$

so that the difference $M_{100}(\pi)-K$ is equal to -0.002403.

The calculation carried out in this paper supply us with 200 partial quotients, and for these I have computed the geometric mean

$$M_{200}(\pi) = 2.727378 .$$

Thus the difference $M_{200}(\pi)-K$ is equal to 0.041828.

Since $M_{200}(\pi)-K$ is more than 17 times as large as the numerical

n	$M_n(\pi)-K$
10	+0.6755
20	−0.1978
30	+0.2010
40	+0.4622
50	+0.3441
60	+0.0191
70	−0.0974
80	+0.1537
90	+0.1317
100	−0.0024
110	−0.0124
120	+0.0328
130	+0.0816
140	+0.1555
150	+0.1685
160	+0.1310
170	+0.0893
180	+0.1268
190	+0.0458
200	+0.0418

Table II

n	$\sqrt[n]{B_n}-L$
10	+0.8296
20	−0.1762
30	+0.2254
40	+0.4562
50	+0.3173
60	+0.0256
70	−0.0959
80	+0.1728
90	+0.1378
100	−0.0066
110	−0.0338
120	+0.0122
130	+0.0573
140	+0.1194
150	+0.1421
160	+0.1052
170	+0.0532
180	+0.0883
190	+0.0172
200	+0.0163

Table III

value of $M_{100}(\pi)-K$, one is compelled to draw the conclusion that the fine agreement between $M_{100}(\pi)$ and K is accidental. That this is actually so is confirmed by table II, where the value of $M_n(\pi)-K$ is given for $n=10, 20, \ldots, 200$. The numerically smallest value in this table is precisely $M_{100}(\pi)-K$. This fact is illustrated in fig. 1, where the values $M_n(\pi)-K$ are plotted as ordinates against the values of n as abscissas.

Khintchine [8] and Lévy [9, 10] have proved independently of each other that $\lim_{n\to\infty} \sqrt[n]{B_n}$ for almost all numbers is an absolute constant, and Lévy has found its value to

$$L = e^{\frac{\pi^2}{12\log 2}} = 3.275823\,.$$

In table III, I have computed the values of $\sqrt[n]{B_n}-L$, where the B_n's are taken from the convergents in the continued fraction expansion of π. The values in table III agree with those in table II as far as the signs are concerned, and the numerically smallest value occurs also in table III for $n=100$.

REFERENCES

[1] J. Wallis: *Tractatus de algebra.* London 1685.

[2] J. H. Lambert: *Beiträge zum Gebrauch der Mathematik und deren Anwendung,* zweiter Teil. Berlin 1770.

[3] O. Perron: *Die Lehre von den Kettenbrüchen.* Leipzig 1929.

[4] D. H. Lehmer: *Euclids algorithm for large numbers.* Amer. Math. Monthly 45 (1938), pp. 227–233.

[5] D. H. Lehmer: *Note on an absolute constant of Khintchine.* Amer. Math. Monthly 46 (1939), pp. 148–152.

[6] P. Pedersen: *Kædebrøksudvikling for π.* Mat. Tidsskrift B (1945), pp. 142–144.

[7] A. Khintchine: *Metrische Kettenbruchprobleme.* Compositio Math. 1 (1935), pp. 361–382.

[8] A. Khintchine: *Zur metrischen Kettenbruchtheorie.* Compositio Math. 3 (1936), pp. 276–285.

[9] P. Lévy: *Sur les lois de probabilité dont dépendent les quotients complets et incomplets d'une fraction continue.* Bull. Soc. Math. France 57 (1929), pp. 178–194.

[10] P. Lévy: *Sur le développement en fraction continue d'un nombre choisi au hasard.* Compositio Math. 3 (1936), pp. 286–303.

OM LÖSNINGEN AV ETT ÖVERBESTÄMT EKVATIONSSYSTEM ENLIGT MINSTA KVADRATMETODEN

OLLE PERSSON

Följande uppsats utgör en sammanställning av förut kända och inom utjämningsräkningen tillämpade metoder, sedda från statistisk synvinkel och med någon matematisk motivering.

Inom statistiken ställes man ofta inför problemet att på grundval av observationer av en stokastisk variabel (t. ex. mätvärden) uppskatta okända konstanter som ingå i denna variabels frekvensfunktion, vilken antages given till sin analytiska form. En metod enligt vilken sådana uppskattningar kunna erhållas är den s. k. maximum-likelihood-metoden. Vid användandet av denna bildas ett uttryck L för sannolikhetstätheten i den simultana fördelningen för de observerade värdena. Genom att maximera L erhållas uppskattningar av de okända konstanterna. Maximeringen av L överför det ursprungliga problemet till problemet att lösa ett ekvationssystem, vilket, om ett stort antal parametrar ingå i fördelningen, kommer att innehålla ett stort antal ekvationer. I allmänhet är detta system icke linjärt. Ett annat problem som uppstår är att bestämma noggrannheten (medelfelen) i de erhållna uppskattningarna.

I fortsättningen skall ett speciellt problem av ovan angivet slag studeras.

Problem: *Det förligger en serie oberoende mätningar* $l_1, l_2, \ldots, l_r$, *där* l_i *antages vara normalfördelad med medelvärdet* λ_i *och medelavvikelsen* σ. *Vidare antages att*

$$\lambda_i = \varphi_i(x_1, \ldots, x_s), \quad i = 1, 2, \ldots, r \quad (s < r)\,, \tag{1}$$

där $x_1, \ldots, x_s$ *äro okända konstanter. Bilda uppskattningar av konstanterna* $x_1, \ldots, x_s$ *på grundval av mätningarna* l_i.

Likelihood-funktionen L blir i detta fall

$$L = \frac{1}{\left(\sqrt{2\pi}\,\sigma\right)^r} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{r}(l_i-\lambda_i)^2} .$$

$x_1, \ldots, x_s$ skola bestämmas så att L blir maximum d. v. s. så att

$$\sum_{i=1}^{r} (l_i - \lambda_i)^2 \tag{2}$$

blir minimum (minsta kvadratmetoden).

Problemet är i allmänhet svårt, men om funktionerna φ_i äro *linjära* bli räkningarna genomförbara utan några förhandstips om konstanterna x_i.

Fortsättningsvis göres följande inskränkning:

$$\varphi_i(x_1, \ldots, x_s) = a_{i1}x_1 + \ldots + a_{is}x_s \, .$$

Införas storheterna $v_i = \lambda_i - l_i$ kan systemet (1) skrivas

$$a_{i1}x_1 + \ldots + a_{is}x_s = l_i + v_i \quad (i = 1, 2, \ldots, r)$$

och uttrycket (2):

$$\sum_{i=1}^{r} v_i^2 \, .$$

För den fortsatta framställningen förutsättes kännedom om elementära räkneoperationer med matriser.

Införas nu följande matrisbeteckningar

$$\begin{bmatrix} a_{11} & \ldots & a_{1s} \\ \vdots & & \vdots \\ a_{r1} & \ldots & a_{rs} \end{bmatrix} = \underset{rs}{\boldsymbol{A}}, \quad \begin{bmatrix} x_1 \\ \vdots \\ x_s \end{bmatrix} = \underset{s1}{\boldsymbol{X}}, \quad \begin{bmatrix} l_1 \\ \vdots \\ l_r \end{bmatrix} = \underset{r1}{\boldsymbol{L}}, \quad \begin{bmatrix} v_1 \\ \vdots \\ v_r \end{bmatrix} = \underset{r1}{\boldsymbol{V}},$$

övergår (1) och (2) till

$$\underset{rs}{\boldsymbol{A}}\,\underset{s1}{\boldsymbol{X}} = \underset{r1}{\boldsymbol{L}} + \underset{r1}{\boldsymbol{V}} \quad (\underset{rs}{\boldsymbol{A}} \text{ förutsättes ha rangen } s) \tag{1}$$

$$\underset{1r}{\boldsymbol{V}^*}\,\underset{r1}{\boldsymbol{V}} \, . \tag{2}$$

I fortsättningen slopas dimensionsbeteckningarna.

Minimum av $q = \boldsymbol{V}^*\boldsymbol{V}$ sökes:

$$q = \boldsymbol{V}^*\boldsymbol{V} = (\boldsymbol{X}^*\boldsymbol{A}^* - \boldsymbol{L}^*)(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{L}) = \boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} - \boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{L} - \boldsymbol{L}^*\boldsymbol{A}\boldsymbol{X} + \boldsymbol{L}^*\boldsymbol{L} \, .$$

Men

$$\boldsymbol{L}^*\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{L} \quad (\boldsymbol{L}^*\boldsymbol{A}\boldsymbol{X} \text{ är en skalär}) \, ,$$

och därför blir

$$q = \boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} - 2\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{L} + \boldsymbol{L}^*\boldsymbol{L} \, .$$

Bildas nu den totala differentialen av q erhålles

$$\begin{aligned} dq &= d\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}d\boldsymbol{X} - 2d\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{L} \\ &= 2d\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} - 2d\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{L} = 2d\boldsymbol{X}^*(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}^*\boldsymbol{L})\,. \end{aligned}$$

Faktorn $2(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}^*\boldsymbol{L})$ utgör de partiella derivatorna av q med avseende på $x_1, \ldots, x_s$. Lösningen till systemet

$$\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} = \boldsymbol{A}^*\boldsymbol{L} \tag{3}$$

leder alltså till maximum eller minimum av q. Efter ytterligare en differentiering av q erhålles

$$d^2q = 2d^2\boldsymbol{X}^*(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}^*\boldsymbol{L}) + 2d\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}d\boldsymbol{X}\,.$$

På grund av (3) försvinner första termen och

$$d^2q = 2(\boldsymbol{A}d\boldsymbol{X})^*(\boldsymbol{A}d\boldsymbol{X}) = 2\|\boldsymbol{A}d\boldsymbol{X}\|^2 \geqq 0\,.$$

Likhetstecknet kan emellertid aldrig gälla, ty då vore $\boldsymbol{A}d\boldsymbol{X} = \boldsymbol{0}$, vilket strider mot förutsättningen om rangen hos $\boldsymbol{A}$.

Alltså är $d^2q > 0$ och systemets (3) lösning leder till minimum för q. Det ursprungliga problemet är således överfört till problemet att lösa det symmetriska ekvationssystemet (3).

Choleskys metod *för upplösning av linjära symmetriska ekvationssystem*: För lösandet av större linjära ekvationssystem av den typ som erhålles enligt minsta kvadratmetoden ha ett flertal metoder utarbetats. Av dessa torde Choleskys trianguleringsmetod vara mest lämpad för handmaskinräkning. Den medför bland annat minimalt skrivarbete till skillnad från andra liknande metoder och dessutom erhållas medelfelen i uppskattningarna av de okända konstanterna på ett mycket överskådligt och enkelt sätt.

Antag inledningsvis att det är möjligt att genomföra en sönderläggning av $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{S}$ enligt följande identitet:

$$\boldsymbol{T}^*\boldsymbol{T} = \boldsymbol{S}\,,$$

$$\boldsymbol{T} = \begin{bmatrix} t_{11} & \cdots & t_{1s} \\ \vdots & \ddots & \vdots \\ & & t_{ss} \end{bmatrix}$$ (triangulär matris; alla element till vänster om huvuddiagonalen äro nollor).

Då erhållas följande räkneregler (för enkelhets skull genomföres kalkylen med en matris av ordningen 3):

$$\begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{bmatrix} = \begin{bmatrix} t_{11} & 0 & 0 \\ t_{12} & t_{22} & 0 \\ t_{13} & t_{23} & t_{33} \end{bmatrix} \cdot \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{23} \\ 0 & 0 & t_{33} \end{bmatrix}$$

$$
\begin{aligned}
t_{11}^2 &= s_{11}\,, & t_{11} &= \sqrt{s_{11}} \\
t_{11}t_{12} &= s_{12}\,, & t_{12} &= \frac{s_{12}}{t_{11}} \\
t_{11}t_{13} &= s_{13}\,, & t_{13} &= \frac{s_{13}}{t_{11}} \\
t_{12}^2+t_{22}^2 &= s_{22}\,, & t_{22} &= \sqrt{s_{22}-t_{12}^2} \\
t_{12}t_{13}+t_{22}t_{23} &= s_{23}\,, & t_{23} &= \frac{s_{23}-t_{12}t_{13}}{t_{22}} \\
t_{13}^2+t_{23}^2+t_{33}^2 &= s_{33}\,, & t_{33} &= \sqrt{s_{33}-t_{13}^2-t_{23}^2}\,,
\end{aligned}
$$

allmänt för $m \leqq n$:

$$\sum_{i=1}^{m} t_{im}t_{in} = s_{mn}\,,$$

$$t_{mm}t_{mn} = s_{mn} - \sum_{i=1}^{m-1} t_{im}t_{in}\,.$$

Om nu alla diagonalelementen t_{ii} äro reella och $\neq 0$ blir denna kalkyl genomförbar. Det återstår således att visa att så är fallet.

Från algebran förutsättes följande hjälpsats bekant: *Om* $\boldsymbol{X}^*\boldsymbol{S}\boldsymbol{X} > 0$ *för alla* $\boldsymbol{X} \neq \boldsymbol{0}$ *så är varje avsnittsdeterminant*

$$|\boldsymbol{S}_k| = \begin{vmatrix} s_{11} & \ldots & s_{1k} \\ \vdots & & \vdots \\ s_{1k} & \ldots & s_{kk} \end{vmatrix} > 0 \quad (k = 1, \ldots, s)\,.$$

Förutsättningen i denna sats är uppfylld för ett ekvationssystem erhållet enligt minsta kvadratmetoden, vilket framgår av andra differentialen vid minimiundersökningen ($d\boldsymbol{X}^*\boldsymbol{A}^*\boldsymbol{A}d\boldsymbol{X} = d\boldsymbol{X}^*\boldsymbol{S}d\boldsymbol{X} > 0$ för godtyckligt $d\boldsymbol{X}$).

Man ser lätt att

$$\boldsymbol{T}_k^*\boldsymbol{T}_k = \boldsymbol{S}_k \quad (k = 1, \ldots, s)$$

där $\boldsymbol{T}_k$ är den till $\boldsymbol{S}_k$ hörande triangulära matrisen. Alltså är

$$|\boldsymbol{S}_k| = |\boldsymbol{T}_k^*||\boldsymbol{T}_k| = |\boldsymbol{T}_k|^2\,.$$

Vidare gäller att

$$|\boldsymbol{T}_k| = t_{kk}|\boldsymbol{T}_{k-1}|$$

och följaktligen är

$$|\boldsymbol{S}_k| = t_{kk}^2|\boldsymbol{S}_{k-1}|\,.$$

Av hjälpsatsen följer då att t_{kk} kan väljas >0 ($k=1, \ldots, s$) och att också $|\boldsymbol{T}| > 0$.

Systemet

$$\boldsymbol{SX} = \boldsymbol{R} \qquad (\boldsymbol{R} = \boldsymbol{A}^*\boldsymbol{L})$$

kan således skrivas

$$\boldsymbol{T}^*\boldsymbol{TX} = \boldsymbol{R}$$

eller

$$\boldsymbol{TX} = (\boldsymbol{T}^*)^{-1}\boldsymbol{R} = \boldsymbol{C}\,.$$

När $\boldsymbol{C}$ är känd är detta ett *triangulärt* ekvationssystem, som kan lösas successivt med avseende på de obekanta, varvid x_s löses ut först.

Hur skall nu $(\boldsymbol{T}^*)^{-1}\boldsymbol{R}$ beräknas?

När metoden för upplösningen av $\boldsymbol{S}$ härleddes utnyttjades identiteten $\boldsymbol{T}^*\boldsymbol{T}=\boldsymbol{S}$, d. v. s.

$$\boldsymbol{T} = (\boldsymbol{T}^*)^{-1}\boldsymbol{S}\,.$$

Varje kolumn i $\boldsymbol{T}$ erhölls således genom operation av $(\boldsymbol{T}^*)^{-1}$ på motsvarande kolumn i $\boldsymbol{S}$. Om nu en godtycklig kolumn $\boldsymbol{R}$ lägges till höger om $\boldsymbol{S}$ kommer tydligen $(\boldsymbol{T}^*)^{-1}\boldsymbol{R}$ att erhållas på precis samma sätt som $\boldsymbol{T}$-kolumnerna. Alltså blir elementen c_m $(m=1, \ldots, s)$ i kolumnen $\boldsymbol{C}$

$$c_m = \frac{r_m - \sum_{i=1}^{m-1} c_i t_{im}}{t_{mm}}\,.$$

Beräkning av medelfelen *i uppskattningarna av de okända konstanterna*: Föregående resultat kan nu användas även för beräkning av den uppskattning av den okända konstanten σ^2 som erhålles enligt maximum-likelihood-metoden samt medelfelen i uppskattningarna av $x_1, \ldots, x_s$.

Uppskattningen s^2 av σ^2 enligt maximum-likelihood-metoden blir

$$s^2 = \frac{(\boldsymbol{V}^*\boldsymbol{V})_{\min}}{r}\,.$$

Om medelvärdet av s^2 beräknas erhålles

$$\frac{r-s}{r}\,\sigma^2\,,$$

varför uppskattningen

$$s^2 = \frac{(\boldsymbol{V}^*\boldsymbol{V})_{\min}}{r-s}$$

har korrekt medelvärde och bör användas då antalet överbestämningar $r-s$ är litet. I fortsättningen skall denna sista medelvärdeskorrigerade uppskattning användas. Det väsentliga vid beräkningen av s^2 är att beräkna $(\boldsymbol{V}^*\boldsymbol{V})_{\min}$.

Enligt föregående är

$$V^*V = X^*A^*AX - 2L^*AX + L^*L\,.$$

Om uppskattningarna $X = (A^*A)^{-1}A^*L$ insättas erhålles

$$(V^*V)_{\min} = L^*A(A^*A)^{-1}A^*A(A^*A)^{-1}A^*L - 2L^*A(A^*A)^{-1}A^*L + L^*L$$
$$= L^*L - L^*A(A^*A)^{-1}A^*L\,.$$

Nu är

$$(A^*A)^{-1} = (T^*T)^{-1} = T^{-1}(T^*)^{-1}\,,$$

och således blir

$$(V^*V)_{\min} = L^*L - L^*AT^{-1}(T^*)^{-1}A^*L = L^*L - [(T^*)^{-1}A^*L]^*(T^*)^{-1}A^*L\,.$$

Om matrissymbolerna tolkas erhålles

$$(V^*V)_{\min} = \sum_{i=1}^{r} l_i^2 - \sum_{i=1}^{s} c_i^2\,.$$

Medelfelen i uppskattningarna av $x_1, \ldots, x_s$ kunna erhållas om följande sats tillämpas: *Om* $f = f(l_1, \ldots, l_r)$, *där mätningarna* l_i *antagas vara oberoende och normalfördelade med medelvärdena* λ_i *och samma medelavvikelse* σ, *är*

$$\sigma_f^2 = \sigma^2 \sum_{i=1}^{r} \left(\frac{\partial f}{\partial l_i}\right)^2 \quad \textit{(medelfelets fortplantningslag).}$$

En linjär funktion av $x_1, \ldots, x_s$ kan i matrisform skrivas:

$$f = FX$$

$$\left(\text{exempel: } f = x_1 + x_2 + x_3 = [1\ 1\ 1]\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},\quad F = [1\ 1\ 1]\right).$$

För att satsen skall kunna tillämpas måste uppskattningarna av $x_1, \ldots, x_s$ uttryckas i de ursprungliga oberoende mätningarna. Alltså

$$f = F(A^*A)^{-1}A^*L$$

och

$$df = F(A^*A)^{-1}A^*dL\,.$$

Härav erhålles

$$\left[\frac{\partial f}{\partial l_1} \cdots \frac{\partial f}{\partial l_r}\right] = F(A^*A)^{-1}A^*$$

och om uppskattningen s^2 insättes i st. f. σ^2 erhålles uppskattningen s_f^2 av σ_f^2:

$$s_f^2 = s^2F(A^*A)^{-1}A^*A(A^*A)^{-1}F^* = s^2F(A^*A)^{-1}F^*\,.$$

På samma sätt som förut blir

$$s_f^2 = s^2[(T^*)^{-1}F^*]^*(T^*)^{-1}F^*\,,$$

och medelfelet i f erhålles om kolumnen F^* lägges på godtycklig plats till höger om $\boldsymbol{S}$ och upplösningen utföres precis som förut. Till höger om $\boldsymbol{T}$, under $\boldsymbol{F}^*$ återfinnes då kolumnen $(\boldsymbol{T}^*)^{-1}\boldsymbol{F}^*$. Kallas elementen i denna f_i erhålles

$$s_f^2 = s^2 \sum_{i=1}^{s} f_i^2 \,.$$

Speciellt blir f = uppskattningen av x_1 om

$$\boldsymbol{F} = [1\ 0\ \ldots\ 0]$$

och f = uppskattningen av x_2 om

$$\boldsymbol{F} = [0\ 1\ \ldots\ 0] \quad \text{etc.}$$

Vid tillämpningen av minsta kvadratmetoden på praktiska problem kunna räkningarna lämpligen uppställas enligt följande schema:

$\begin{matrix} a_{11} & \ldots & a_{1s} \\ \vdots & & \vdots \\ a_{r1} & \ldots & a_{rs} \end{matrix}$	$\begin{matrix} l_1 \\ \vdots \\ l_r \end{matrix}$	
$\begin{matrix} s_{11} & \ldots & s_{1s} \\ \vdots & & \vdots \\ s_{1s} & \ldots & s_{ss} \end{matrix}$	$\begin{matrix} r_1 \\ \vdots \\ r_s \end{matrix}$	$\begin{matrix} 1 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \end{matrix}$
	$\sum_{i=1}^{r} l_i^2$	
$\begin{matrix} t_{11} & \ldots & t_{1s} \\ & \ddots & \vdots \\ & & t_{ss} \end{matrix}$	$\begin{matrix} c_1 \\ \vdots \\ c_s \end{matrix}$	$\begin{matrix} f_{11} & & \\ \vdots & \ddots & \\ f_{s1} & \ldots & f_{ss} \end{matrix}$
$x_1 \ldots x_s$	$\sum_{i=1}^{r} l_i^2 - \sum_{i=1}^{s} c_i^2$	$\sum_{i=m}^{s} f_{im}^2 \ (m = 1, \ldots, s)$

eller med matrisbeteckningar:

$\boldsymbol{A}$	$\boldsymbol{L}$		
$\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{S}$	$\boldsymbol{A}^*\boldsymbol{L} = \boldsymbol{R}$	$\boldsymbol{E}$	$\boldsymbol{F}^*$
	$\boldsymbol{L}^*\boldsymbol{L}$		
$\boldsymbol{T}$	$(\boldsymbol{T}^*)^{-1}\boldsymbol{R}$	$(\boldsymbol{T}^*)^{-1}$	$(\boldsymbol{T}^*)^{-1}\boldsymbol{F}^*$
$\boldsymbol{X}^*$	$(\boldsymbol{V}^*\boldsymbol{V})_{\min}$	$(\boldsymbol{T}^*\boldsymbol{T})_{ii}^{-1}$	$\boldsymbol{F}(\boldsymbol{T}^*\boldsymbol{T})^{-1}\boldsymbol{F}^*$

Följande exempel må tjäna till att underlätta förståelsen av ovan presenterade metoder.

Exempel: För att uppskatta höjden över en viss nivåyta på tre punkter A, B, C samt medelfelen i dessa uppskattningar utfördes avvägning enligt fig. 1 (stigning i pilens riktning). Punkterna D, E, F ligga vid nivåytan, vars höjd sättes $=0$. Höjdskillnaderna

$$\begin{aligned} l_1 &= 1, & l_4 &= 1, \\ l_2 &= 2, & l_5 &= 2, \\ l_3 &= 3, & l_6 &= 1, \end{aligned}$$

vilka antagas vara normalfördelade med samma medelavvikelse σ, uppmättes.

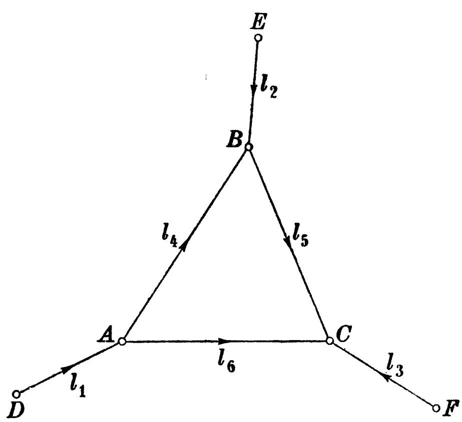

Fig. 1

Om för de okända höjderna A, B, C införas beteckningarna x_1, x_2, x_3 erhålles följande överbestämda ekvationssystem:

$$\begin{array}{rrrl} x_1 & & & \approx l_1 = 1 \\ & x_2 & & \approx l_2 = 2 \\ & & x_3 & \approx l_3 = 3 \\ -x_1 & +x_2 & & \approx l_4 = 1 \\ & -x_2 & +x_3 & \approx l_5 = 2 \\ -x_1 & & +x_3 & \approx l_6 = 1\,, \end{array}$$

och räkneschemat blir

1			1			
	1		2			
		1	3			
-1	1		1			
	-1	1	2			
-1		1	1			
3	-1	-1	-1	1		
-1	3	-1	1		1	
-1	-1	3	6			1
			20			
$\sqrt{3}$	$-\frac{1}{\sqrt{3}}$	$-\frac{1}{\sqrt{3}}$	$-\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$		
	$\sqrt{\frac{8}{3}}$	$-\frac{4}{3}\cdot\sqrt{\frac{3}{8}}$	$\frac{2}{3}\cdot\sqrt{\frac{3}{8}}$	$\frac{1}{3}\cdot\sqrt{\frac{3}{8}}$	$\sqrt{\frac{3}{8}}$	
		$\sqrt{2}$	$6\cdot\frac{1}{\sqrt{2}}$	$\frac{1}{2\sqrt{2}}$	$\frac{1}{2\sqrt{2}}$	$\frac{1}{\sqrt{2}}$
$\frac{5}{4}$	$\frac{7}{4}$	3	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$

Ur räkneschemat erhållas följande uppskattningar:

$$s^2 = \frac{3}{2}\cdot\frac{1}{3} = \frac{1}{2}$$

$$x_1 = \frac{5}{4}, \quad x_2 = \frac{7}{4}, \quad x_3 = 3$$

$$s_{x_1}{}^2 = s_{x_2}{}^2 = s_{x_3}{}^2 = s^2\cdot\frac{1}{2} = \frac{1}{4}.$$

LITTERATUR

N. Arley—K. Rander Buch: *Sandsynlighedsregning.* 4. udg., København 1950.

A. Bjerhammar: *Application of calculus of matrices to method of least squares with special reference to geodetic calculations.* Tekniska Högskolans handlingar nr 49, Stockholm 1951. (Innehåller utförlig litteraturförteckning.)

H. Cramér: *Mathematical methods of statistics.* Uppsala 1945.

R. Zurmühl: *Matrizen.* Berlin 1950.

A MECHANICAL PROBLEM BY H. WHITNEY[1])

ARNE BROMAN

1. The following problem is discussed in Courant-Robbins: *What is mathematics?*[2].

Suppose that a carriage is movable on a horizontal rail (the s axis of fig. 1). A rod can move without friction around a horizontal axle, perpendicular to the rail, through the lower end of the rod. A certain motion

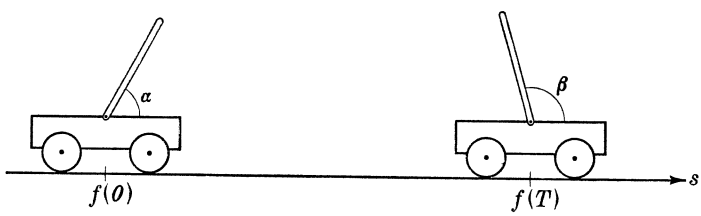

Fig. 1

is prescribed for the carriage: $s=f(t)$, $0 \leqq t \leqq T$, where f is a given function of the time t, s is the position at the time t, and the motion starts at the time $t=0$ and ceases at the time $t=T$. There is given to the rod a certain initial inclination at the time $t=0$, viz. the angle α in fig. 1. It is assumed that if the rod falls down into a horizontal position sometime during the time $0<t<T$, it remains in this position. Is it possible to choose α so that the rod stands vertically at the time $t=T$, i.e. that the angle β of fig. 1 equals $\pi/2$?

Courant-Robbins give the following solution of this problem. The physical assumption is made that the inclination angle β at the time $t=T$ is a continuous function of α, say $\beta=\beta(\alpha)$. Obviously, $\alpha=0$ gives $\beta=0$ and $\alpha=\pi$ gives $\beta=\pi$. By a known theorem on continuous functions there exists a number α_1 between 0 and π such that $\beta(\alpha_1)=\pi/2$. Hence, the answer to the question is yes.

2. However, the physical assumption in the solution of Courant-Robbins seems a bit hazardous[3]. We therefore want to dig a little deeper into the problem. A natural way then is to study the differential equation for the motion of the rod. In this connection we must impose some conditions on the function $s=f(t)$. We choose to assume that it has a continuous second derivative $\ddot{f}(t)$ in the considered interval $0 \leqq t \leqq T$. (We denote differentiation with respect to time by dots.) We further assume that $\dot{f}(0)=\dot{f}(T)=0$, i.e. that the velocity of the carriage is zero at the times $t=0$ and $t=T$, and that the rod has the angular velocity zero at the time $t=0$.

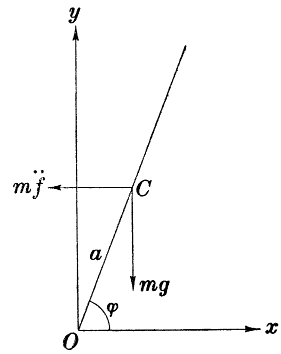

Fig. 2

Fig. 2 gives the situation at the time t. Here O is the lower end of the rod and C its center of gravity. The segment OC is a units of length. The inclination angle of the rod is denoted by φ and its weight by mg, this force being applied at C. The frame of reference xy is in the vertical plane of the s axis of fig. 1; the x axis and s axis have the same direction. The frame xy has a translation motion with respect to the s axis (which is supposed to be at rest in a newtonian frame of reference). We compensate for the motion of the xy frame by applying[4] the fictitious force $m\ddot{f}$ at C (equally directed as the negative x axis if $\ddot{f}$ is positive; otherwise the force is directed as the positive x axis). The moment equation[5] with respect to the point O reads

$$I\ddot{\varphi} = m\ddot{f}a\sin\varphi - mga\cos\varphi\,, \tag{1}$$

where I is the moment of inertia of the rod with respect to O.

Equation (1) is the differential equation for the motion of the rod: I, m, a, and g are given constants, $\ddot{f}$ is a given continuous function of t, and φ is the function of t looked for; when the function $\varphi(t)$ is known, then the motion of the rod is of course also known.

The right-hand member of (1) is a continuous function of φ and t. This function satisfies a Lipschitz condition with respect to φ. A known theorem[6] on differential equations then shows that given initial values of $\varphi(0)$ and $\dot{\varphi}(0)$, there exists a solution of (1) and that this solution is uniquely determined. In the following we denote the solution of (1), which corresponds to the initial values $\varphi(0)=\alpha$ and $\dot{\varphi}(0)=0$ (i.e. at the time $t=0$ the inclination angle of the rod is α and its angular velocity is zero), by $\varphi=\varphi_\alpha(t)$.

Further, a known theorem[7] on differential equations shows that to a given function $\varphi=\varphi_{\alpha_1}(t)$, a given interval $0 \leqq t \leqq T$, and a given number $\varepsilon>0$, there corresponds a number $\eta>0$ such that $|\varphi_\alpha(t)-\varphi_{\alpha_1}(t)|<\varepsilon$ if $|\alpha-\alpha_1|<\eta$ and t belongs to the interval. We express this property in the following by saying that the curve $\varphi = \varphi_\alpha(t)$ is situated in the *ε-channel* around the curve $\varphi = \varphi_{\alpha_1}(t)$ for $|\alpha - \alpha_1| < \eta$ and $0 \leqq t \leqq T$.

3. We now remove for a while the stops for the rod in the horizontal positions so that the rod can revolve all around the point O. Then the final angle β, considered as a function $\beta = \beta(\alpha)$ of the initial angle α, is obviously defined and continuous for all real α. Further, this function satisfies the condition

$$\beta(\alpha + 2\pi) = \beta(\alpha) + 2\pi \tag{2}$$

for every α, for if α is increased by 2π, the motion of the rod is not changed and β is also increased by 2π. Equation (2) shows that the function $\beta = \beta(\alpha)$ is bounded neither upwards nor downwards. Hence the function assumes both positive and negative values. According to the already used theorem on continuous functions there exists at least one number α_1 such that $\beta(\alpha_1) = \pi/2$.

Now the following question arises: Is there also any α_1 such that the curve $\varphi = \varphi_{\alpha_1}(t)$ is situated within the strip $0 < \varphi < \pi$? In connection with this question the curves in fig. 3 should be studied: the lower curve gives a solution for our problem, the upper curve does not, because φ assumes the value π and the rod then falls into a horizontal position and remains there.

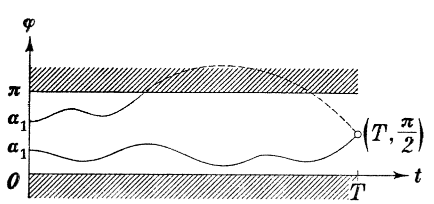

Fig. 3

4. We now restrict ourselves to the class[8] K of the connected parts of the curves $\varphi = \varphi_\alpha(t)$ which have their left-hand end point on the side AB of the rectangle $ABCD$ of fig. 4 and their right-hand end point on any of the other sides. By equation (1) and our assumptions on the function f there is a number δ, $0 < \delta < \pi/2$, such that $\ddot{\varphi} < 0$ if $0 < \varphi \leqq \delta$ and $\ddot{\varphi} > 0$ if $\pi - \delta \leqq \varphi < \pi$. Hence, every curve in the class K which has entered the lower narrow rectangle of fig. 4 is "sucked down" to AD, and an analogous property holds for the curves of K which have entered the upper narrow rectangle.

Suppose that no curve of the class K passes through the middle point M of CD. We then divide K into two classes, K_1 and K_2, such that the curves in K_1 have their right-hand end point on AD or DM and the curves in K_2 their right-hand end point on BC or CM. We further divide the side AB into two point sets, E_1 and E_2, such that E_1 consists of the left-hand end points of the curves in K_1 and E_2 of the left-hand end points of the curves in K_2. None of these sets is empty: e.g. E_1 contains the whole interval $(0, \delta)$ and E_2 the whole interval $(\pi-\delta, \pi)$. Because the segment AB is a connected set of points there exists a point α_1 belonging to one of E_1 and E_2, e.g. E_1, which is a limit point of E_2. There is accordingly in E_2 a sequence of points $\{\alpha_n\}_2^\infty$ such that $\lim_{n\to\infty} \alpha_n = \alpha_1$. Hence every ε-channel of the curve $\varphi=\varphi_{\alpha_1}(t)$ contains a curve of K_2, and the curve $\varphi=\varphi_{\alpha_1}(t)$ has a positive distance to the rectangle side AD in fig. 4 (otherwise some curves of K_2 would be sucked down to AD;

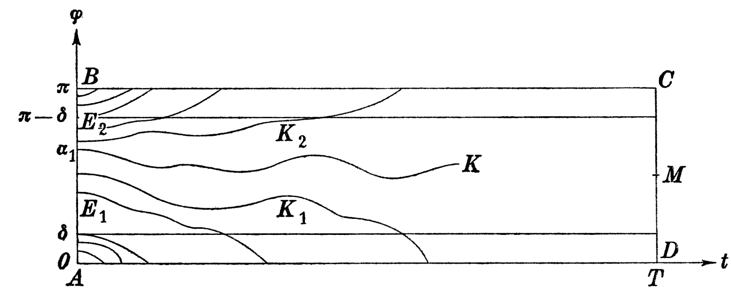

Fig. 4

this is contrary to the definition of the class K_2). It has further a positive distance to BC, because it belongs to the class K_1. Then there is a number $\varepsilon>0$ such that the corresponding ε-channel of the curve $\varphi=\varphi_{\alpha_1}(t)$ contains no point of AD and BC and does not contain M. In this ε-channel there are curves of K_2. This, however, contradicts the definition of the class K_2. Hence, the class K contains at least one curve passing through M. Our mechanical problem is thereby solved[9]: the anwer is yes.

5. If the motion of the rod is described for $0 \leqq t \leqq T$ by the function $\varphi=\varphi_{\alpha_1}(t)$ of section **4**, then the rod probably has a positive kinetic energy at the time $t=T$ and it will fall down into a horizontal position at some later time (we suppose that the carriage stands still for $t \geqq T$). We can then ask if α can be chosen so that the rod never falls down.

To give an answer to this question we make the additional assumption[10] that $\ddot{f}(T)=0$. Then the right-hand member of (1) is a continuous function of φ and t for $0 \leqq t < \infty$, satisfying a Lipschitz condition. Hence, by the problem already solved, there exists a sequence $\{\alpha_n\}_1^\infty$ of values of α such that the curve $\varphi=\varphi_{\alpha_n}(t)$ lies in the strip $0<\varphi<\pi$ for $0 \leqq t \leqq nT$ and passes through the point $(nT, \pi/2)$. This sequence has at least one limit point α_∞. From the assumption that $\varphi=\varphi_{\alpha_\infty}(t)$ takes the value 0 or π

for some $t>0$ there is easily deduced a contradiction, using an ε-channel around the curve. Hence there is a value α_∞ of α such that the rod never falls down. It can be shown that only the following two cases can occur: 1) $\varphi_{\alpha_\infty}(t)=\pi/2$ for $t \geqq T$, 2) $\varphi_{\alpha_\infty}(t)$ approaches $\pi/2$ asymptotically as t tends to ∞.

NOTES

[1] Lecture given at the 13th Scandinavian Mathematical Congress in Helsingfors in August 1957.

[2] COURANT-ROBBINS: *What is mathematics?* (New York 1941, 3rd impr. 1946), pp. 319–321. According to these authors the problem was proposed by H. Whitney.

[3] In saying this we do not intend to criticize the interesting and well-written book by Courant and Robbins. They have included the problem to get an application of the mentioned theorem on continuous functions.

[4] See e.g. J. NIELSEN: *Lærebog i Rationel Mekanik, II* (3rd ed., København 1952), pp. 27–29, where this fictitious force is studied for the motion of a particle. It is left as an exercise to the reader to prove that the fictitious forces on all parts of the rod can be replaced by the force $m\ddot{f}$ applied at C.

[5] See e.g. J. NIELSEN, *loc. cit.*, p. 209, the italicized theorem.

[6] See e.g. CODDINGTON-LEVINSON: *Theory of ordinary differential equations* (New York 1955, International series in pure and applied mathematics), chapt. 1, sect. 6.

[7] See e.g. CODDINGTON-LEVINSON, *loc. cit.*, p. 22, theorem 7.1.

[8] Can two curves of the class K have a point in common? The answer to this question is probably no, but the author has not found a proof. If this were proved, the rest of the proof in section **4** could be considerably simplified.

[9] An interesting alternative solution has been proposed to the author by Dr. M. Tideman. We give a concise account of his proof. — Denote the right-hand member of (1) by $h(\varphi, t)$. Consider the differential equation

$$I\ddot{\varphi} = h^*(\varphi, t)\,, \tag{3}$$

where $h^*=h$ if $0 \leqq \varphi \leqq \pi$, $h^*=h(0,t)$ if $\varphi \leqq 0$, and $h^*=h(\pi,t)$ if $\varphi \geqq \pi$. It is readily verified that if a solution curve of (3) leaves the rectangle $ABCD$ of fig. 4, it can never enter it again. As above in section **2** it is seen that to the initial values $\varphi(0)=\alpha$ $(0<\alpha<\pi)$ and $\dot{\varphi}(0)=0$ there corresponds a unique solution. Denote this solution by $\varphi=\varphi_\alpha(t)$. It is seen that $\varphi_\alpha(T)$ is a continuous function of α, that $\varphi_\alpha(T)<\delta$ for $\alpha \leqq \delta$, and that $\varphi_\alpha(T)>\pi-\delta$ for $\alpha \geqq \pi-\delta$, where δ has the same meaning as in fig. 4. Hence there exists a number α_1, $0<\alpha_1<\pi$, such that $\varphi_{\alpha_1}(T)=\pi/2$. The curve $\varphi=\varphi_{\alpha_1}(t)$ has no point in common with AD or BC. It then gives the answer yes to the question.

[10] It is possible to get rid of this assumption. We refrain from this, however, in order to avoid a more complicated proof.

EN BEMÆRKNING OM TO RÆKKER

DAVID FOG

I en artikel i NMT[1] har Viggo Brun fremdraget de to klassiske rækker

$$\ln 2 = 1-\tfrac{1}{2}+\tfrac{1}{3}-\tfrac{1}{4}+\ \ldots \tag{1}$$

og

$$\frac{\pi}{4} = 1-\tfrac{1}{3}+\tfrac{1}{5}-\tfrac{1}{7}+\ldots\ . \tag{2}$$

Det følgende skal blot vise en simpel forbindelse mellem disse rækker; de kan nemlig betragtes som specialtilfælde af en og samme formel.

Som udgangspunkt tages ligningen

$$\int \mathrm{tg}^{p-1}x\,dx = \frac{\mathrm{tg}^p x}{p} - \int \mathrm{tg}^{p+1}x\,dx, \qquad p \geqq 1\ , \tag{3}$$

hvis rigtighed umiddelbart kan indses ved differentiation. Af (3) fås

$$\int_0^{\frac{\pi}{4}} \mathrm{tg}^{p-1}x\,dx = \frac{1}{p} - \int_0^{\frac{\pi}{4}} \mathrm{tg}^{p+1}x\,dx\ , \tag{4}$$

som ved gentagen anvendelse giver

$$\int_0^{\frac{\pi}{4}} \mathrm{tg}^{p-1}x\,dx = \frac{1}{p} - \frac{1}{p+2} + \ldots + \frac{(-1)^q}{p+2q} + (-1)^{q+1}\int_0^{\frac{\pi}{4}} \mathrm{tg}^{p+2q+1}x\,dx\ . \tag{5}$$

Da $\mathrm{tg}^n x$ for $n \to \infty$ konvergerer ligeligt mod 0 i ethvert lukket delinterval til intervallet $0 \leqq x < \frac{\pi}{4}$, vil restleddet i (5) gå mod 0 for $q \to \infty$, og vi har

$$\int_0^{\frac{\pi}{4}} \mathrm{tg}^{p-1}x\,dx = \frac{1}{p} - \frac{1}{p+2} + \frac{1}{p+4} - \frac{1}{p+6} + \ldots\ . \tag{6}$$

For $p=1$ er denne formel identisk med (2), medens den for $p=2$ giver

$$\tfrac{1}{2}\ln 2 = \tfrac{1}{2}-\tfrac{1}{4}+\tfrac{1}{6}-\tfrac{1}{8}+\ldots\ ,$$

der er ensbetydende med (1).

---

[1] Viggo Brun: *On the problem of partitioning the circle so as to visualize Leibniz' formula for π.* NMT 3 (1955), s. 159–166.

EINE NOTIZ N. H. ABELS FÜR A. L. CRELLE AUF EINEM MANUSKRIPT OTTO AUBERTS

KURT-R. BIERMANN und VIGGO BRUN

Nach dem Tode August Leopold Crelles (1780–1855), des Berliner Freundes und Protektors N. H. Abels, kamen die Manuskripte für das von Crelle 1826 begründete »Journal für die reine und angewandte Mathematik«, dessen 197. Band z. Zt. im Erscheinen begriffen ist, in den Besitz der Berliner Buchhandlung Asher. 1867 beschloß die Berliner Akademie der Wissenschaften auf Antrag von Weierstrass, Borchardt, Kummer und Kronecker, diese Manuskripte anzukaufen. Schon damals begann der später verwirklichte Plan sich abzuzeichnen, eine Gesamtausgabe der Werke C. G. J. Jacobis, Steiners und Lejeune Dirichlets herauszugeben. Dabei sollte, wie Weierstrass es in der Vorrede zum 1. Band von C. G. J. Jacobis Ges. Werken (Berlin 1881) ausführt, jede einzelne Arbeit »vor dem Abdruck einer sorgfältigen Revision unterworfen und nicht nur von Schreib- und Druckfehlern, sondern auch von sonstigen, offenbar bloß durch Versehen entstandenen Unrichtigkeiten möglichst gereinigt, im Übrigen aber der ursprüngliche Text als historisches Dokument treu beibehalten werden«. Es versteht sich, daß bei der Durchführung dieses Programms die Manuskripte der in Crelles Journal erschienenen Abhandlungen für die Revision benutzt worden sind, worauf Weierstrass auch a. a. O. ausdrücklich hinweist.

Die im erwähnten Nachlaß Crelles befindlichen Abhandlungen N. H. Abels wurden von der Berliner Akademie an L. Sylow und Sophus Lie (eine Empfangsbestätigung vom 22. 11. 1873 wird im Archiv der Deutschen Akademie der Wissenschaften aufbewahrt) zur Benutzung bei der Herausgabe der Werke Abels ausgeliehen, und zwar die Originale der Nrn. 11, 16, 19, 20, 21 des I. und der Nr. 23 des II. Bandes der Ausgabe Holmboes. Später ist auf Veranlassung von Weierstrass die Leihgabe in ein Geschenk umgewandelt worden, damit die Manuskripte zusammen mit dem übrigen Nachlaß Abels in Oslo aufbewahrt werden.

Heute befinden sich noch unter den im Archiv der Deutschen Akademie der Wissenschaften zu Berlin liegenden Handschriften aus Crelles Nachlaß folgende zwei Abhandlungen Abels:

> Beweis eines Ausdrucks, von welchem die Binomial-Formel ein einzelner Fall ist (Crelles Journal, 1, Bln. 1826, S. 287–292)

und

Fernere mathematische Bruchstücke aus N. H. Abels Briefen. Schreiben Abels an Legendre (Crelles Journal, 6, Bln. 1830, S. 73–80).

In die Ausgabe Sylow–Lie ist die französische Fassung dieser beiden Artikel aufgenommen worden.

Bei der Durchsicht und Ordnung des Nachlasses Crelles im Archiv der Deutschen Akademie der Wissenschaften wurde festgestellt, daß sich auf dem Manuskript von O. G. D. Aubert, das später in Crelles Journal, 5, Bln. 1830, S. 163–173 unter dem Titel »Bemerkungen zu den Aufgaben und Lehrsätzen S. 96, 97, 98 im ersten Heft zweiten Bandes dieses Journals« in erweiterter Form gedruckt wurde, ein unveröffentlichter Nachsatz von Abels Hand befindet[1]:

»Liebster Hr. Geheimrath! Schon vor 14 Tage oder wohl noch mehr sandte ich ihnen die Fortsetzung der Abhandlung über die elliptischen Functionen, die Sie wohl erhalten haben werden. und ist es mein Wunsch daß Sie damit zufrieden waren. — Die vorstehende Abhandlung ist von einem hiesigen noch sehr jungen Mathematiker; der aber gewiß Talente hat. Er beschäftiget sich wie Steiner vorzüglich mit geometrischen Sachen. Es sollte mich freuen wenn Sie diese Abhandlung zur Aufnahme in das Journal würdig finden. Auch zweifle ich nicht daran. — Und er wird darinn gewiß Aufmunterung finden. — Eben in diesem Augenblick habe ich endlich das zweite Heft bekommen. Ich danke ihnen dafür recht sehr. Das dritte und vierte aber noch nicht, allein sie werden wohl bald folgen. — Nachdem ich meine Abhandlung fortgeschickt hatte, fand ich noch einen algemeinen Satz über die elliptischen Functionen. Ich bitte Sie daher noch folgendes der »Addition au memoire prècédent« hin zufügen:

On peut encore trouver la valeur de l'angle ψ de beaucoup d'autres manières: Ainsi par exemple on aura sans ambigüité:

$$\begin{aligned}\psi = \theta &+ 2\,\mathrm{Arctang}\left\{\mathrm{tang}\,\theta\cdot\sqrt{1-e^2\sin^2\theta^{\mathrm{II}}}\right\} + \\ &+ 2\,\mathrm{Arctang}\left\{\mathrm{tang}\,\theta\cdot\sqrt{1-e^2\sin^2\theta^{\mathrm{IV}}}\right\} + \cdots + \\ &+ 2\,\mathrm{Arctang}\left\{\mathrm{tang}\,\theta\cdot\sqrt{1-e^2\sin^2\theta^{(2n)}}\right\}\end{aligned}$$

Mit Hochachtung
ihr ergebenster
N. Abel«

---

[1] An fremden Sprachen hat Abel fast ausschließlich die französische benutzt; Deutsch beherrschte er weniger gut. Hier ist sein deutsch geschriebener Brief ohne sprachliche Verbesserungen wiedergegeben.

Die Notiz Abels auf dem Manuskript Otto Auberts trägt kein Datum, ist aber höchstwahrscheinlich Ende Februar oder Anfang März 1828 geschrieben, da er sich darin auf die Fortsetzung der Arbeit »Recherches sur les fonctions elliptiques« mit dem Zusatz »Addition au mémoire précédent« bezieht. Abel hat diesen zweiten Teil am 12. Februar 1828 an Crelle gesandt.

Otto Aubert (1809–1838), ursprünglich Pädagoge an der Christiania Katedralskole, war später als Lehrer der norwegischen Sprache und der mathematischen Wissenschaften für die beiden Söhne des schwedisch-norwegischen Kronprinzen Oscar, d. h. für die späteren Könige Karl XV und Oscar II, angestellt. Ausser der Arbeit im Crelleschen Journal hat Aubert im Programm für die Christiania Katedralskole für das Jahr 1833 eine geometrische Studie, »Echantillon d'une analyse spherique«, publiziert. Diese nach dem Tode Abels ausgearbeitete Abhandlung dürfte die Meinung Abels, daß er »gewiß Talente hat«, vollauf bestätigen. Wie Abel ist auch er an Tuberkulose gestorben.

Aubert fuhr 1834 nach Deutschland. Seine dort geschriebenen Briefe sind teilweise in »Landflygtige« von Sofie Aubert Lindbæk (Kristiania 1910) veröffentlicht worden. Er hat darin Crelle, J. Steiner (1796–1863), J. Plücker (1801–1868) und A. F. Möbius (1790–1868) sehr unterhaltsam geschildert.

BOKMELDINGER

MAX HIEKE: *Vektoralgebra.* (Mathematisch-naturwissenschaftliche Bibliothek 4.) B. G. Teubner Verlagsgesellschaft, Leipzig, 1956. 6+154 S., 67 Fig. Geb. DM 9.20.

(Innholdsfortegnelse i NMT 5 (1957), s. 110.)

Denne boken gir først en grundig innføring i den elementære vektoralgebraen (til sine tider vel pirkete etter anmelderens smak). Som vanlig i tysk litteratur benyttes gotiske bokstaver for vektorer, og på figurene brukes skrevne gotiske bokstaver, noe som vel virker forvirrende på de fleste skandinaviske lesere til å begynne med. Som vanlig i eldre tysk litteratur benyttes parentes for å betegne skalart produkt og klammerparentes for vektorprodukt, mens vel det mest vanlige og mer anvendelige i dag er $\cdot$ og $\times$.

En steiler uvilkårlig når en på s. 3 finner følgende kategoriske utsagn: »Die Kraft ist kein Vektor.« Som forklaring på dette anføres at utenom vektoregenskapene kommer det også an på angrepspunktet når det gjelder virkningen på et fast legeme. Hieke lar således ikke vektorbegrepet omfatte linjevektorer. Etter anmelderens mening er dette uheldig. Mengden av alle like linjevektorer langs en bestemt virkningslinje danner jo en undermengde av mengden av alle vektorer med samme retning og tallverdi. I sin språkbruk klarer da ikke Hieke alltid å unngå å benytte betegnelsen vektor om en linjevektor. Skulle en forfølge Hiekes tanke konsekvent, burde en jo heller ikke kalle en posisjonsvektor for en vektor.

Etter den elementære vektoralgebraen følger et inngående kapitel om linjevektorer med anvendelser. Anmelderen måtte her lete lenge før han fant definisjonen på ekvivalente vektorsystemer, som står svært bortgjemt.

I neste kapitel behandles lineære transformasjoner ved hjelp av matriser. Etter anmelderens mening er det naturligere å benytte dyader i vektoranalysen. Brukes summasjonskonvensjonen blir det vel så enkelt med dyader, og dessuten har dyade-begrepet store fordeler når en skal innføre tensorbegrepet.

I siste kapitel som er kalt tensorer behandles kartesiske tensorer av 2. orden uten at summasjonskonvensjonen innføres. På grunn av dette blir ligningene ofte unødig kompliserte og uoversiktlige. Nytt for anmelderen i en bok av denne type var en rekke anvendelser på krystallografi. — Boken inneholder ingen øvelsesoppgaver.

I forordet sies at det også skal komme et bind om vektoranalyse.

Oddvar Bjørgum

M. V. Wilkes: *Automatic digital computers.* Methuen & Co., London, 1956. 10+305 pp. sh. 42/-.

(Innholdsfortegnelse i NMT 5 (1957), s. 47.)

George R. Stibitz — Jules A. Larrivee: *Mathematics and computers.* McGraw-Hill Book Co., New York, Toronto, London, 1957. 7+228 pp. sh. 37/6.

(Innholdsfortegnelse i NMT 5 (1957), s. 155.)

Når man i årene like etter krigen skulle sette seg inn i teorien for og bruken av de nye elektroniske regnemaskiner, var nesten den eneste mulighet å besøke personlig de forskjellige institutter hvor slike maskiner var i bruk eller under konstruksjon. Noen representativ litteratur av betydning fantes ikke. Den eksplosive utvikling av disse maskiner i løpet av det siste tiår har imidlertid etterhvert også satt spor i litteraturen. Tekniske spesialartikler utkommer nå i et omfang som gjør det vanskelig å følge med i alt som skrives, og større bøker for begynnere og viderekomne er det også blitt skrevet endel av.

Anmelderen får ofte spørsmål etter egnet litteratur om elektroniske regnemaskiner. I et så nytt felt har det ikke utkrystallisert seg noen »klassiske« bøker, men jeg tror det kan være nyttig å innlede denne anmeldelse med en liste over noen av de viktigste oversiktsverker innen feltet.

De grunnleggende Princeton-rapporter av Burks, Goldstine og von Neumann fra 1946–48 er vel nå ikke å oppdrive. Av annen »eldre« litteratur kan nevnes

D. R. Hartree: *Calculating instruments and machines.* Univ. of Illinois Press, 1949; Cambridge Univ. Press, 1950.

Denne boken behandler flere forskjellige typer av regneinstrumenter og maskiner. Spesiallitteratur om siffermaskiner kommer først fra 1950 og utover:

Engineering Research Associates: *High-speed computing devices.* Mc Graw-Hill, New York 1950.

H. Rutishauser — A. Speiser — E. Stiefel: *Programmgesteuerte digitale Rechengeräte.* Birkhäuser, Basel 1951.

Synthesis of electronic computing and control circuits. Ann. Comp. Lab., Harvard Univ., v. 27, 1951.

M. V. Wilkes — D. J. Wheeler — S. Gill: *The preparation of programs for an electronic digital computer.* Addison-Wesley, Cambridge (Mass.) 1951.

A. D. Booth — K. H. V. Booth: *Automatic digital calculators.* Butterworth, London 1953.

R. K. Richards: *Arithmetic operations in digital computers.* Van Nostrand, New York 1955.

Til denne rekke av »seriøse« verker slutter seg den nye bok av Wilkes som her skal anmeldes. Forfatteren har stått i spissen for konstruksjonen og driften av maskinen EDSAC i Cambridge, vel den første operative maskin i Europa. Det er ikke urimelig at Wilkes i sin nye bok øser av sine rike erfaringer med denne maskin. En viss skjevhet i fremstillingen er det kanskje blitt av den grunn, f. eks. ved den sterke vekt på serie- kontra parallell-maskiner. Men stort sett blir eksemplene fra EDSAC balansert av en fyldig omtale også av andre maskin-typer.

Første kapitel er en historisk oversikt. Som alle engelskmenn er Wilkes meget stolt av Charles Babbage, som i begynnelsen av forrige århundre hadde komplette planer for en virkelig helautomatisk regnemaskin — selvsagt på mekanisk basis.

Annet kapitel gir prinsippene for maskinenes virkemåte, de forskjellige tall-representasjoner og litt om koding. Det siste utdypes i neste kapitel, hvor Wilkes bruker EDSAC-koden som eksempel. Prinsippene for subrutiner behandles nokså inngående; på dette felt er jo Wilkes og hans stab banebrytere.

Et kapitel om relé-maskiner faller litt utenfor rammen, men er i seg selv meget morsomt. Deretter følger en oversikt over de forskjellige former for »hukommelse«. Dette kapitel forutsetter en del kjennskap til fysikk for å kunne fordøyes fullstendig.

Et kapitel om logikken i elektroniske regnekretser er meget godt, med en innføring i bruken av Boolesk algebra i slike kretser. Etter anmelderens mening er imidlertid Wilkes' utvalg av representative konstruksjoner litt tilfeldig.

I siste kapitel kommer forfatterens erfaringer helt til sin rett. Han gir en meget klar analyse av de problemer som reiser seg ved konstruksjon og drift av en stor elektronisk maskin. I et appendiks avlives myten om

»tenkende maskiner«, og boken avsluttes med en fyldig litteraturfortegnelse og et stikkord-register.

Wilkes' nye bok er ikke lett lesning for begynnere, men den vil gi noe igjen for arbeidet. Også viderekomne kan finne en rekke ting av interesse i boken, spesielt i forfatterens kloke vurderinger av fordeler og mangler ved de forskjellige systemer. Boken kan trygt anbefales; etter anmelderens mening er den noe av det beste som er skrevet om elektroniske regnemaskiner.

I tillegg til alvorlige bøker om de nye regnemaskiner kommer det også en voksende rekke av »lettvektere«, som varierer fra en sterkt populariserende innføring til ren »science fiction«. Den annen bok som her skal anmeldes, av Stibitz og Larrivee, hører nærmest til den sterkt populariserende type. Ihvertfall en av forfatterne, Stibitz, har et ganske kjent navn innen kretsen av fagfolk (men det vil irritere mange lesere at han benytter anledningen til å fremheve sin egen innsats).

Boken behandler ikke bare siffermaskiner, men også analogmaskiner. Ifølge forordet skal den gi legmannen en viss forståelse av den nye regneteknikk som er vokset frem. Som bakgrunn gis et »lynkurs« i høyere matematikk. Forfatterne fremhever at en matematiker godt kan hoppe over dette avsnitt. Hvis han allikevel leser det, vil han støte på følgende gullkorn: De reelle tall inndeles i tre grupper, rasjonale, irrasjonale *og* transcendente!

Etter at et spinkelt matematisk grunnlag er lagt, omtales endel problemer som løses på automatiske maskiner, samt disse maskiners konstruksjon og virkemåte i korte trekk. Det sier seg selv at forfatterne ikke kan gå i detalj i en bok som denne, men man savner en omtale av den siste utvikling på området. Litt antikvert virker også den stadige presisering av forskjellen mellom »rene« og »anvendte« matematikere; ifølge forfatterne er det bare de siste som befatter seg med elektroniske regnemaskiner.

Boken er skrevet i en lett og humoristisk stil. Illustrasjonen av en binær addisjonsenhet ved hjelp av trappevendere er f. eks. litt av en perle. Men av og til blir forfatterne vel bråkjekke, eller hva skal man si til følgende: Etter en (meget kort) innføring av den deriverte følger en (enda kortere) forklaring på det bestemte integral, under titelen »Reconstructing the crime«!

Hvert kapitel avsluttes med noen forslag til litteratur. Etter anmelderens mening er disse bøker og artikler nokså tilfeldig utvalgt. Derimot er litteraturfortegnelsen i slutten av boken meget fyldig.

Ernst S. Selmer

OPPGAVER TIL LØSNING

Løsninger av oppgavene 141–144 sendes til oppgaveredaktøren, professor R. Tambs Lyche, Matematisk Institutt, Blindern, Oslo. Slike løsninger vil bli trykt i et følgende hefte i den utstrekning plassen tillater, dog vanligvis bare den beste løsning av hver oppgave. Løsninger av oppgaver i dette hefte må være sendt innen 15. august 1958.

De øvrige oppgaver i dette hefte er enklere, og løsninger af dem vil ikke bli trykt.

Redaksjonen er takknemlig for forslag til oppgaver. Slike forslag kan sendes til oppgaveredaktøren, helst sammen med forslagsstillerens egen løsning.

141. La n være et naturlig tall og $P_{2n+1}(x)$ Legendres polynom av orden $2n+1$. Beregn integralet

$$\int_0^1 \frac{P_{2n+1}(x)}{x(1+x)^{2n+2}}\,dx\,.$$

W. Ljunggren

142. La m og p være naturlige tall. Bevis at

$$\Delta_n = \begin{vmatrix} \binom{m}{p} & \binom{m+1}{p+1} & \cdots & \binom{m+n-1}{p+n-1} \\ \binom{m+1}{p-1} & \binom{m+2}{p} & \cdots & \binom{m+n}{p+n-2} \\ \cdots & \cdots & \cdots & \cdots \\ \binom{m+n-1}{p-n+1} & \binom{m+n}{p-n+2} & \cdots & \binom{m+2n-2}{p} \end{vmatrix} = \prod_{i=0}^{n-1} \frac{\binom{m+i}{p-i}\binom{m+p+2i}{i}}{\binom{p+i}{i}\binom{p}{i}}\,.$$

Johs. Lohne

143. Lad $\varphi(x)$ betegne antallet af naturlige tal $\leqq x$ med given tværsum t. Bevis at

$$\lim_{x\to\infty} \frac{\varphi(x)}{(\log_{10} x)^t} = (t!)^{-1}\,.$$

Ove J. Munch

144. Lad E være en ellipse med halvakser a og b, $a \geqq b \geqq 0$. Idet $P(z)$ er et polynomium af n'te grad med hovedkoefficient 1, skal man bevise $(n \geqq 1)$:

$$\max_{z \in E} |P(z)| \geqq \left(\frac{a+b}{2}\right)^n + \left(\frac{a-b}{2}\right)^n .$$

For hvilket polynomium indtræffer lighed?

Ove J. Munch

145. Binomialkoefficienterne tænkes opstillet på sædvanlig måde i Pascal's trekant. Bestem tre på hinanden følgende binomialkoefficienter (i samme række), således at den midterste er middeltal mellem de to andre.
Eks.: $\binom{14}{4} = 1001$, $\binom{14}{5} = 2002$, $\binom{14}{6} = 3003$.

David Fog

146. Om ellipsen $b^2x^2 + a^2y^2 = a^2b^2$ omskrives et rektangel der det ene par parallelle sider har vinkelkoeffisienten k. Bestem rektangelets areal og finn det største av disse rektangler.

147. Vis at den største femkant som har gitte sider a, b, c, d og e må kunne innskrives i en sirkel.

LØSNINGER

137. Bevis formelen

$$\sum_{i=1}^{n} (-1)^{i-1} \frac{\binom{a}{i}\binom{b}{i}}{i\binom{n}{i}} = \sigma_a + \sigma_b - \sigma_n ,$$

der a, b og n er slike naturlige tall at $a+b \geqq n \geqq b \geqq a$, og

$$\sigma_r = 1 + \frac{1}{2} + \ldots + \frac{1}{r} .$$

W. Ljunggren

(I den opprinnelige oppgave var eksponenten for -1 feilaktig angitt som i.)

Løsning: Vi setter

$$f(a, b, n) = \sum_{i=1}^{n} (-1)^{i-1} \frac{\binom{a}{i}\binom{b}{i}}{i\binom{n}{i}}, \quad g(a, b, n) = \sigma_a + \sigma_b - \sigma_n ,$$

for $a+b \geqq n \geqq b \geqq a > 0$. Identiteten

$$\frac{\binom{a+1}{i}}{\binom{n+1}{i}} = \frac{n-a}{n+1} \cdot \frac{\binom{a+1}{i}}{\binom{n}{i}} + \frac{a+1}{n+1} \cdot \frac{\binom{a}{i}}{\binom{n}{i}}$$

gir ved multiplikasjon med $(-1)^{i-1} \binom{b}{i} i^{-1}$ og summasjon fra 1 til n:

$$f(a+1, b, n+1) = \frac{n-a}{n+1} f(a+1, b, n) + \frac{a+1}{n+1} f(a, b, n) ,$$

når alle argumenter ligger i f's definisjonsområde. En enkel regning viser at også

$$\frac{n-a}{n+1} g(a+1, b, n) + \frac{a+1}{n+1} g(a, b, n) = g(a+1, b, n+1) .$$

Det er da nok å vise at $f(a, b, b) = g(a, b, b)$ og $f(1, b, b+1) = g(1, b, b+1)$. Den siste likheten innses umiddelbart; den første bevises ved induksjon etter a, idet

$$f(a, b, b) = \sum_{i=1}^{a} (-1)^{i-1} \frac{\binom{a}{i}}{i} = -\int_0^{-1} \sum_{i=1}^{a} \binom{a}{i} x^{i-1} dx$$

$$= -\int_0^{-1} \frac{(1+x)^a - 1}{x} dx = -\int_0^{-1} \frac{(1+x)(1+x)^{a-1} - 1}{x} dx$$

$$= f(a-1, b, b) - \int_0^{-1} (1+x)^{a-1} dx = f(a-1, b, b) + \frac{1}{a} .$$

Helge Tverberg

Også løst av L. Carlitz.

139. Låt p vara ett primtal. Bevisa kongruensen

$$\binom{m}{p^n} \equiv \left[\frac{m}{p^n}\right] \pmod{p} .$$

Bernt Lindström

Løsning: It is familiar that if

$$m = a_0 + a_1 p + a_2 p^2 + \dots \quad (0 \leqq a_i < p) ,$$
$$k = b_0 + b_1 p + b_2 p^2 + \dots \quad (0 \leqq b_i < p) ,$$

then

(1) $$\binom{m}{k} \equiv \binom{a_0}{b_0} \binom{a_1}{b_1} \binom{a_2}{b_2} \dots \pmod{p} .$$

In particular

$$\binom{m}{p^n} \equiv a_n \quad (\mathrm{mod}\, p)\,.$$

Since

$$\left[\frac{m}{p^n}\right] = a_n + a_{n+1}p + \dots,$$

it follows that

$$\binom{m}{p^n} \equiv \left[\frac{m}{p^n}\right] \quad (\mathrm{mod}\, p)\,.$$

L. Carlitz

Også løst av C. U. Jensen og Helge Tverberg, som begge løser oppgaven uten å forutsette kongruensen (1) kjent.

140. La d_n være determinanten

$$d_n = \begin{vmatrix} \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n} \end{vmatrix},$$

der n er et naturlig tall. Vis at

$$d_n = 1 \Big/ \prod_{k=1}^{n} k \binom{n}{k} \binom{n+k}{k}\,.$$

R. Tambs Lyche

Løsning: La a_{ij} være $(x_i + y_j)^{-1}$. Determinanten $|a_{ij}|$ blir da av formen

$$\frac{P(x_1, \dots, x_n, y_1, \dots, y_n)}{\prod_{i=1}^{n} \prod_{j=1}^{n} (x_i + y_j)\,,},$$

der P er et polynom. Er $x_k = x_l$ eller $y_k = y_l$ for noe par $k \neq l$, blir $|a_{ij}| = 0$ fordi to rekker eller søyler blir like. Altså har P som divisor

$$\prod_{i<j\leq n} (x_j - x_i) \cdot \prod_{i<j\leq n} (y_j - y_i)\,.$$

Betraktning av et spesielt ledd, f. eks. ${x_n}^{n-1}{y_n}^{n-1}$, viser at P nettopp er lik det angitte produkt. Ved å sette $x_i = i$, $y_j = j$, får en da umiddelbart oppgavens formel.

Helge Tverberg

Også løst av J. E. Fjeldstad, Johs. Lohne og Asmus L. Schmidt.

SUMMARY IN ENGLISH

PEDER PEDERSEN: *On the expansion of π in a regular continued fraction.* (English.)

In two papers from 1938 and 1939, D. H. Lehmer computed the partial quotients a_n in the continued fraction for π up to a_{100} (cf. references p. 68). Without knowing Lehmer's work, the author in 1945 performed the same computation up to a_{96}. In the present paper, the expansion is continued up to a_{200}, using the step-by-step method developed by Lehmer.

The partial quotients up to a_{200} are given in table I p. 62, and the convergents A_{n-1}/B_{n-1} and A_n/B_n found at the end of each step are listed on pp. 63–65. Two different checks of the final convergents, corresponding to $n = 200$, are performed.

As was also done by Lehmer, the author uses the convergents for a verification of Khintchine's and Lévy's constants. The approximations obtained for $n = 10$, 20, ..., 200 are shown in tables II–III p. 67, and table II is also illustrated graphically p. 66. The unusually close approximation obtained by Lehmer for $n = 100$ must be accidental.

OLLE PERSSON: *On the solution of an over-determined system of equations by the method of least squares.* (Swedish.)

The following problem is considered: Given a set of independent measurements $l_1, l_2, \ldots, l_r$, where l_i is normally distributed with mean value λ_i and standard deviation σ. It is assumed that

$$\lambda_i = a_{i1}x_1 + \ldots + a_{is}x_s, \quad i = 1, 2, \ldots, r \ (s < r),$$

where $x_1, \ldots, x_s$ are unknown constants, which should be estimated.

It is shown that this problem, by the maximum-likelihood-method, leads to the solution of a symmetric system of equations:

$$\boldsymbol{A^*AX} = \boldsymbol{A^*L}$$

(where $\boldsymbol{X}$ and $\boldsymbol{L}$ are column matrices). The method of Cholesky for solving such a system (including the determination of mean errors for the x_i) is described and illustrated by a simple example.

ARNE BROMAN: *A mechanical problem by H. Whitney.* (English.)

A carriage is movable on a horizontal rail (fig. 1 p. 78) with a prescribed motion. A rod can move without friction around a horizontal axle, perpendicular to the rail, through the lower end of the rod. Is it possible to choose the initial inclination α of the rod so that the rod stands vertically when the motion ceases?

The problem, proposed by H. Whitney and discussed in Courant—Robbins: What is mathematics?, is solved (confirmatively) under the assumption of a continuous acceleration of the carriage. One can also choose the initial inclination so that the rod never falls down.

DAVID FOG: *A remark on two series.* (Danish.)

The classical series for $\pi/4$ and for $\ln 2$ are particular cases, with $p=1$ and $p=2$ respectively, of the expansion

$$\int_0^{\frac{\pi}{4}} \operatorname{tg}^{p-1} x\, dx = \frac{1}{p} - \frac{1}{p+2} + \frac{1}{p+4} - \frac{1}{p+6} + \dots, \quad p \geqq 1 .$$

KURT-R. BIERMANN and VIGGO BRUN: *A note by N. H. Abel to A. L. Crelle on a manuscript by Otto Aubert.* (German.)

The manuscripts, left by Crelle, for the "Journal für die reine und angewandte Mathematik", are now deposited with the Deutsche Akademie der Wissenschaften in Berlin. Among the manuscripts is a paper by O. G. D. Aubert: "Bemerkungen zu den Aufgaben und Lehrsätzen S. 96, 97, 98 im ersten Heft zweiten Bandes dieses Journals" (published in Vol. 5 (1830), pp. 163–173). It has an unpublished postscript to Crelle by Abel, introducing Aubert and suggesting an addendum to Abel's "Recherches sur les fonctions elliptiques".