RENDICONTI
del
SEMINARIO MATEMATICO
della
UNIVERSITÀ DI PADOVA

A. DE MATTEIS
B. FALESCHINI

Some arithmetical theorems on base conversions

Rendiconti del Seminario Matematico della Università di Padova, tome 43 (1970), p. 261-268

<http://www.numdam.org/item?id=RSMUP_1970__43__261_0>

SOME ARITHMETICAL THEOREMS ON BASE CONVERSIONS

A. De Matteis - B. Faleschini *)

Abstract - It is shown that the necessary and sufficient condition to represent numbers, given in the N_1-scalte to n_1 significant digits, in a machine with base N_2 and n_2 significant rounded digits, such that inverse conversion from the N_2-scale yields the same n_1 rounded digits in the N_1-scale, is $N_1^{n_1} \leq \lambda N_2^{n_2-1}$. The factor λ is determined for all possible cases.

1. Introduction.

Recently Goldberg [1] has shown by a numerical example that, although $10^8 < 2^{27}$, 27 significant binary digits are not always sufficient to represent accurately decimal numbers with 8 significant digits. He has pointed out the interval [9000000.0, 9999999.9] containing 10^7 numbers with 8 decimal digits, where a binary machine with 27 significant digits has only 8.10^6 numbers. Therefore we cannot make distinct decimal numbers correspond to distinct binary numbers. In this note we shall examine the necessary and sufficient conditions to represent numbers in a machine with a given accuracy.

We shall refer in the following to the numbers in the two scales of notation as to normalized floating-point numbers of different machines, each machine being characterized by the pair (N, n) of the base $N > 1$ sand of the number n of digits for the mantissa; we shall set no restriction on the exponent and, moreover, it will be sufficient to consider positive, non-zero numbers.

Converting a number x of the machine (N_1, n_1) to the N_2-scale and

*) Indirizzo degli A: Centro di Calcolo del C.N.E.N., Via Mazzini, 2 - CA.P. 40138, Bologna.

rounding the result to n_2 significant digits we get a number y of the machine (N_2, n_2), which we call the correspondent of x. Let z be the result rounded to n_1 digits in the N_1-scale of the inverse conversion of y from the N_2-scale. We say that the machine (N_2, n_2) represents with the accuracy of n_1 digits in the N_1-scale (or, briefly, with the accuracy (N_1, n_1)) the numbers of the machine (N_1, n_1) if, for every x, $z=x$. Goldberg has found for the two bases $N_1=10$, $N_2=2$ the sufficient condition $10^{n_1}<2^{n_2-1}$. If $n_1=8$ the smallest integer satisfying this inequality is $n_2=28$. Moreover, as shown by the numerical example above, this is also the number of digits strictly sufficient to represent accurately 8 decimals.

It has been proved [2] that when N_1 and N_2 are not powers of the same integer also the inverse of the theorem established by Goldberg holds; i.e. we have in the machine (N_2, n_2) the accuracy (N_1, n_1) if, and only if, $N_1^{n_1}<N_2^{n_2-1}$. We shall give in this note an alternative proof of this theorem.

When the two bases are powers of the same integer, they may always be reduced to the form $N_1=b^{k_1}$, $N_2=b^{k_2}$, with k_1 and k_2 relatively prime. For this case we prove here that the necessary and sufficient conditions is

$$N_1^{n_1}\leq bN_2^{n_2-1}.$$

For example, two octal digits are necessary to represent accurately floating-point numbers with three binary digits.

The relation between the definition of accuracy given above and the distance between machine numbers is not obvious and it is not sufficient, in general, to verify that in a given interval one machine has more numbers than the other in order to decide on the accuracy of the representation. Consider, as a simple example, the two machines (2, 1) and (10, 1). In the interval between 1 and 2^{28} the binary machine has only 27 numbers, while the decimal one has many more numbers. This notwithstanding, one decimal digit is not sufficient to represent accurately a binary digit: in fact the number 2^{27} converted and rounded to one decimal becomes 10^8 and this last number is reconverted to 2^{26}. Therefore, we shall start by pointing out the connection between accuracy and distances of machine numbers.

2. Accuracy and distances between machine numbers.

Every normalized floating-point number $x>0$ of the machine (N, n) is a number of the form $0\cdot x_1, x_2 \dots x_n \cdot N^{p(x)}$ with $x_1 \neq 0$, $p(x)$ being an integer for which we shall set no restriction. When x is a power of the base N, say N^k, k being an integer, we shall write simply $x=N^k$ instead of $x=\frac{1}{N}N^{k+1}$; however $p(N^k)=k+1$. We shall denote by $\bar{x}$ and x' the machine number predecessor and successor of x, respectively, and by $d(x)=N^{p(x)-n}$ the distance of x from x'. It will be $d(\bar{x})=d(x)$ if $x\neq N^k$, and $d(\bar{x})=\frac{d(x)}{N}$ if $x=N^k$. Adopting the usual rounding procedure, the number x will represent in the machine (N, n) the real numbers of the half-open interval

$$\left[x-\frac{d(x)}{2},\ x+\frac{d(x)}{2}\right[$$

if $x\neq N^k$, and the real numbers of the interval

$$\left[x-\frac{d(x)}{2N},\ x+\frac{d(x)}{2}\right[$$

if $x=N^k$.

The following Lemma shall relate the accuracy to the distances between machine numbers. Comparing two machines, (N_1, n_1) and (N_2, n_2), we will always denote by x the numbers of the first and by y the numbers of the second. For simplicity we will also write $d(x)$ and $d(y)$ instead of $d_1(x)$ and $d_2(y)$; analogously for the exponents $p(x)$ and $p(y)$.

Lemma 1. Let $x>0$ be any normalized floating-point number of the machine (N_1, n_1) and y its correspondent in the machine (N_2, n_2). If

$$d(\bar{y})<d(x), \text{ whenever } x<y,$$

$$d(y)\leq d(\bar{x}), \text{ whenever } x>y,$$

then the machine (N_2, n_2) represents all numbers of (N_1, n_1) with the accuracy (N_1, n_1).

PROOF. 1) $x<y$. The number x belongs to the interval of real numbers represented by y in (N_2, n_2) and therefore, since $x<y$, it will be $y-x\leq\frac{d(\bar{y})}{2}$. The inverse conversion of y gives x back if $y-x<\frac{d(x)}{2}$. This last inequality will be satisfied by the hypothesis $d(\bar{y})<d(x)$.

2) $x>y$. Analogously, $x-y<\frac{d(y)}{2}$ and y reconverts to x if $x-y\leq\frac{d(\bar{x})}{2}$, etc., completing the proof.

For the case $x=y$ it is not necessary to require any condition.

3. Conditions for a given accuracy.

THEOREM 1. Let N_1 and N_2 be not powers of the same integer. The machine (N_2, n_2) represents with the accuracy (N_1, n_1) the floating-point numbers $x>0$ of the machine (N_1, n_1) if, and only if,

$$N_1^{n_1}<N_2^{n_2-1}. \tag{1}$$

PROOF. Condition (1) is sufficient. Let $x>0$ be a number of (N_1, n_1) and y its correspondent. Consider the case $x<y$ with $y\neq N_2^k$. By Lemma 1 there will be in (N_2, n_2) the required accuracy if $d(\bar{y})=d(y)<d(x)$; i.e., if

$$\frac{N_2^{p(y)}}{N_1^{p(x)}}<\frac{N_2^{n_2}}{N_1^{n_1}}.$$

But, in general

$$0\cdot x_1x_2 \ \dots\ x_{n_1}\cdot N_1^{p(x)}=(0\cdot y_1y_2\ \dots\ y_{n_2}+\varepsilon)\cdot N_2^{p(y)}$$

where, if $y\neq N_2^k$, $-N_2^{-n_2}/2\leq\varepsilon<N_2^{-n_2}/2$ (if $y=N_2^k$ the lower bound must be divided by N_2). Since the two mantissas are normalized and moreover, by hypothesis, $x<y$, i.e. $\varepsilon<0$, it follows:

$$\frac{N_2^{p(y)}}{N_1^{p(x)}}=\frac{0\cdot x_1x_2\ \dots\ x_{n_1}}{0\cdot y_1y_2\ \dots\ y_{n_2}+\varepsilon}<\frac{1-N_1^{-n_1}}{N_2^{-1}-N_2^{-n_2}}.$$

Therefore there will be the required accuracy if

$$\frac{1-N_1^{-n_1}}{N_2^{-1}-N_2^{-n_2}}<\frac{N_2^{n_2}}{N_1^{n_1}}$$

which is the condition (1). If $x<y$ and $y=N^k$, then $d(\bar{y})=\frac{d(y)}{N_2}$ and a less restrictive condition than (1) is found.

The proof for the case $x>y$ is analogous and is carried out by distinguishing the case $x\neq N_1{}^k$ from the case $x=N_1{}^k$ (i.e. $0\cdot x_1 \dots x_{n_1}=N_1^{-1}$).

Condition (1) is necessary. We will show that if $N_1^{n_1}>N_2^{n_2-1}$, it is always possible to determine a number x of the machine (N_1, n_1) whose correspondent y reconverts to $z\neq x$. For example, if we can find a number x such that

$$(2) \qquad y-\bar{x}<x-y<y'-x,$$

then $y<x$ (second inequality) and y is closer to $\bar{x}$ than to x (first inequality), so that $z=\bar{x}\neq x$.

Since, for integers p and q, $x=N_1{}^p$ and $y=N_2{}^q$ are respectively numbers of (N_1, n_1) and (N_2, n_2), we shall determine p and q in order to satisfy (2). For the particular values chosen for x and y we have $\bar{x}=N_1{}^p-N_1^{p-n_1}$ and $y'=N_2{}^q+N_2^{q+1-n_2}$. By substitution in (2), we obtain

$$(3) \qquad \alpha N_2{}^q<N_1{}^p<\beta N_2{}^q,$$

where $\alpha=2/(2-N_1^{-n_1})$ and $\beta=(2+N_2^{1-n_2})/2$. Having supposed $N_1^{n_1}>$ $>N_2^{n_2-1}$, then $1<\alpha<\beta$. Taking the logarithms in the base N_1, (3) may be rewritten

$$b_1<p-aq<b_2$$

where $a=\log_{N_1} N_2$, $b_1=\log_{N_1}\alpha$, $b_2=\log_{N_1}\beta$.

Develop now the number a into a continued fraction, and consider the odd convergents to a, P_{2i+1}/Q_{2i+1}, so that $P_{2i+1}-aQ_{2i+1}>0$. Let $p=k\,P_{2i+1}$ and $q=k\,Q_{2i+1}+h$, where k and h are integers to be determined so that

$$b_1<k(P_{2i+1}-aQ_{2i+1})-ha<b_2\,,$$

i.e., so that

$$\frac{b_1+ha}{P_{2i+1}-aQ_{2i+1}}<k<\frac{b_2+ha}{P_{2i+1}-aQ_{2i+1}}\,.$$

Since by hypothesis N_1 and N_2 are not powers of the same integer, the number a is irrational, thus the difference $P_{2i+1}-aQ_{2i+1}$ is different from zero and can be made smaller than any preassigned quantity. Therefore for each h it is sufficient that $\frac{b_2-b_1}{P_{2i+1}-aQ_{2i+1}}>1$ to find an integer k and hence two integers p and q satisfying (3). The proof of the theorem is thus complete.

To show, for example, that one ternary digit is not sufficient to represent one binary digit, let $h=1$. The second convergent to $a=\log 3/\log 2$, $P_1/Q_1=2/1$, gives $k=5$ and hence $p=10$, $q=6$. In fact, the number $x=2^{10}$ converts to $y=3^6$ and the inverse conversion of y gives $z=2^9$.

Let us consider now two bases powers of the same integer, say 2 and 8. As we know, three binary digits are equivalent to one octal digit, but one octal digit is not always sufficient to represent numbers with three binary digits: in fact between 1 and 2 the machine $(N_1, n_1)=$ $=(2, 3)$ has three numbers while the octal machine has none.

For the proof of the next theorem we need to know the floating-point representation of any integer power b^h in a machine with the base $N=b^k$. To this purpose, let q be the quotient of the division of h by k and r the remainder, i.e.

$$h=qk+r \text{ with } 0\leq r<k.$$

Then

$$b^h=\frac{b^r}{N}N^{q+1}.$$

From the above identity we conclude that b^h is exactly representable in every machine with base $N=b^k$, by one of the N-ary digits 1, b, b^2, ..., b^{k-1}; moreover the power of the base N in the normalized representation is $p(b^h)=q+1$.

THEOREM 2. Let $N_1 = b^{k_1}$ and $N_2 = b^{k_2}$, with k_1 and k_2 relatively prime. The machine (N_2, n_2) represents with the accuracy (N_1, n_1) the floating-point numbers $x > 0$ of the machine (N_1, n_1) if, and only if,

$$N_1^{n_1} \leq b N_2^{n_2-1}. \tag{4}$$

PROOF. Consider the numbers of the machine (N_1, n_1) limited by two successive integer powers of b, say b^h and b^{h+1}. These two bounds are exactly represented both in (N_1, n_1) and in (N_2, n_2). Moreover this interval cannot contain either a power of N_1 or of N_2: this means that the numbers in both the machines are equidistant. The ratio of these two distances is a power of b, thus the machine with more numbers in the interval considered represents exactly the numbers of the other one. We conclude that a necessary and sufficient condition for the accuracy required is that the distance between the numbers of (N_2, n_2) be not greater than that of (N_1, n_1). Since

$$b^h = b^{r_1} \cdot N_1^{q_1} = b^{r_2} \cdot N_2^{q_2} \quad \text{where } h = k_1 q_1 + r_1 = k_2 q_2 + r_2\,,$$

we must have

$$N_2^{q_2+1-n_2} \leq N_1^{q_1+1-n_1}$$

which is the same as $k_2(q_2 + 1 - n_2) \leq k_1(q_1 + 1 - n_1)$, or

$$k_2 n_2 \geq k_1 n_1 + k_2 - k_1 + r_1 - r_2\,. \tag{5}$$

We must now determine $\max\limits_h (r_1 - r_2)$, where

$$r_1 \equiv h \bmod k_1$$

$$r_2 \equiv h \bmod k_2\,.$$

Since k_1 and k_2 are relatively prime, it is easy to find that $\max\limits_h (r_1 - r_2) = k_1 - 1$; by substitution in (5) we get

$$k_2 n_2 \geq k_1 n_1 + k_2 - 1$$

and hence (4). Since by intervals of the type $[b^h, b^{h+1}]$ we cover all the range of both machines, the proof is complete.

For example, let $(N_1, n_1)=(2, 27)$ and $N_2=16$. To represent accurately 27 binary digits we need 8 hexadecimal digits. To show that 7 hexadecimal digits are not sufficient, consider the interval between 1 and 2: the binary machine has in this interval 2^{27-1} numbers (including one of the two bounds) and the hexadecimal machine has 16^{7-1} numbers. Therefore in the hexadecimal machine 3.2^{24} numbers are missing in the interval considered.

ACKNOWLEDGEMENT. The authors wish to thank Prof. U. Richard for useful suggestions in the preparation of this paper.

REFERENCES

[1] GOLDBERG I. B.: *27 bits are not enough for 8-digit accuracy,* Comm. ACM **10**, (February 1967), 105-106.

[2] MATULA D. W.: *In-and-out conversions,* Comm. ACM **11**, 1 (January 1968), 47-50.

Manoscritto pervenuto in redazione il 21 ottobre 1969.