MATHEMATICS OF COMPUTATION
VOLUME 42, NUMBER 166
APRIL 1984, PAGES 623-635

A Complete Axiomatization of Computer Arithmetic

By Richard Mansfield

Abstract. We define an axiom system for rounded arithmetic to be complete if we can recover from any model of the axioms the exact algebra from whence it came. A complete set of axioms is given for rounded addition and multiplication.

In some circles of numerical analysis, there is a deep suspicion that some of the scientific and engineering calculations being done by computer do not correspond to external reality. One cause of this unease is that computer arithmetic itself can be inexact. For example, we all know that $((a + b)^2 - (a^2 + 2ab))/b^2$ is one, yet, in IBM single precision with $a = 100$ and $b = .01$, the result is around 39. In addition to this unavoidable consequence of rounding, it is also a sad fact that not all floating-point packages are free of mistakes. The addition routines on both the UNIVAC mainframe and the TRS-80 Color Computer (and probably many more) contain serious bugs. It is depressing that commerical incentives do not exist for manufacturers and firmware suppliers to improve their product.

One method for dealing with some of these problems is to establish mathematical definitions of just what floating-point arithmetic should do. The IEEE has proposed standards incorporating such definitions, thus providing the computer industry with a scale against which it can measure its efforts. It is also hoped that, by formulating mathematical laws concerning computer arithmetic, we can produce a theory which will allow a greater degree of justified confidence in actual computations. The theory has not as yet progressed to the stage of having such useful implications.

Knuth [1, p. 214] has given some laws for the algebra of computer arithmetic. Kulisch and Miranker [2] have carried out the project in more detail. Kulisch and Miranker also give efficient algorithms conforming to the IEEE standards. Neither of these sources consider the problem of giving a complete axiomatization.

There are two causes of inaccuracy in computer arithmetic, rounding and overflow. In this preliminary research, I propose to simplify the problem by pretending that overflow does not exist. One justification for this is that overflow generally occurs rarely and when it does happen, the programmer is usually informed with a warning message. Rounding, on the other hand, occurs silently and invidiously on nearly every step of a numerical calculation. Future research will deal with combining the two sources of error. Within this limitation we will produce a complete axiomatization of rounded floating-point arithmetic. Of course, if any of this work is to be useful, the axioms should be simple enough for each comprehension. I am

Received April 19, 1983; revised June 14, 1983.
1980 *Mathematics Subject Classification*. Primary 68B99, 68C99, 08A99, 03C99, 65G05.

afraid that this goal has not yet been achieved. We are instead bound by Einstein's maxim, "Be as simple as possible, but no simpler."

Let R be an ordered field. A rounding function on R is a monotone projection, i.e., a function O mapping R into itself such that

$$\begin{aligned}&O(O(x)) = O(x),\\ &x \leqslant y \quad \text{implies} \quad O(x) \leqslant O(y),\\ &O(0) = 0\colon \quad O(1) = 1\colon \quad O(-x) = -O(x).\end{aligned}$$

Let S be the range of O and let $+ +$ and $* *$ be the field operations of addition and multiplication. We can define corresponding operations on S by means of the equations,

$$x + y = O(x + + y), \qquad x * y = O(x * * y),$$

for all x and y in S. The field R also bequeaths its order relation to S.

The goal of this research is to give a complete set of axioms for the ordered algebra S. In logic the word complete has two standard meanings. In one sense, an axiom system is complete if it either proves or refutes every formula in a given language. This is not what is meant here. In our sense, an axiom set is complete for a given concept if it completely defines the concept as, for instance, the group axioms define the concept group. Some concepts such as "finite" or "Archimedian" are not axiomatizable. Our axioms will be complete in the sense that any model S for the theory can be extended to an ordered field with a rounding function as defined above. In other words, the above paragraphs define a model class, the class of rounded algebras, and we propose to axiomatize that class. Actually we start with a method for reconstructing the field from the algebra and then invent axioms to prove theorems we already know to be true.

Our decision to sweep overflow under the rug determines our choice of the standard model for the theory. The intended model is to let S be the set of real numbers expressible with a mantissa of some fixed finite precision but an arbitrary integer exponent, e.g., all reals of the form $m * 2^n$ where n is an arbitrary integer but m is an integer of absolute value less than 2^{24}. The rounding function O is the standard rounding to the nearest screen point with the proviso that a number exactly half way between two screen points will not be hung up like the famous donkey but will round to the nearest odd mantissa. This condition is given in order to simplify some of our axioms. Using the convention, "When in doubt round up" would better conform to actual hardware but would further complicate an already complicated axiom set without requiring any new ideas. It is commonly believed that IBM uses truncation rather than rounding, but I am convinced that this is not so, at least for addition and subtraction. If hex 8 is used for the padded guard digit, IBM's method gives ordinary rounding. The role of the standard model in the ensuing theory is the central one of guarding the consistency of the axioms. All the axioms are meant to be true of this model. Of course this can easily be considered more of a hope than a guarantee. So far all the false axioms have proved to be correctable, some with more difficulty than others. In any list of over forty axioms there is bound to be some errors.

Our plan is to first introduce axioms for the reconstruction of the additive group. Then further axioms will be introduced for multiplication and division. Our first group of axioms is essentially that of ordered groups without associativity. Remember that in all that follows $+$ and $*$ are the operations on S, and that these axioms refer to S.

1. $x + y = y + x$,
2. $x \leqslant y$ implies $x + z \leqslant y + z$,
3. $x \leqslant y$ implies $-y \leqslant -x$,
4. $x \leqslant y \leqslant z$ implies $x \leqslant z$,
5. $x \leqslant y \leqslant x$ implies $x = y$,
6. $x + 0 = x$,
7. $x + (-x) = 0$.

A crucial fact about the standard model is that there is a function $r_+(x, y)$ mapping S into S such that

$$r_+(x, y) + + (x + y) = x + + y.$$

In fact Knuth [1, Theorem B, p. 220] proves that $r_+(x, y)$ is definable in the standard model. Let $x' = (x + y) - y$ and let $y' = (x + y) - x'$. Then Knuth proves that

$$r_+(x, y) = (x - x') + (y - y').$$

Note the lack of symmetry in the definition. In spite of this r_+ must be commutative since both $+$ and $+ +$ are commutative. We shall use the notation $r(x + y)$ for $r_+(x, y)$, i.e., in the expression $r(x + y)$, we consider x, y, and $+$ to be separate variables. The main task before us is to axiomatize this function without reference to any structure external to S.

Definition.

$x \lll y$ iff $x + y = y$,
$x \ll y$ if $x \lll y$ or
($y \neq 0$ and there is a z such that ($r(y + z) = 0$ and $x \lll z$)).

In the standard model, $x \lll y$ pretty much means that the most significant digit of x is at least 25 places less significant than the most significant digit of y. The exceptions occur when either x or y is a power of 2. The relation $x \ll y$ means that even though $x + y$ may be unequal to y, the sum $x + y$ does not change any digits of y, it just appends more digits to the least significant end of y. In other words, the nonzero digits of x and y do not overlap. This is similar to the relation defined by Kulisch and Miranker [2, p. 292].

Here are some more axioms.

8. $r(x + y) = r(y + x)$,
9. $x \ll y \ll z$ implies $x \ll z$,
10. $x \lll y \lll z$ implies $x \lll z$,
11. $r(x + y) \lll x + y$,
12. $0 < y < z$ and $x \lll y$ implies $x \lll z$,
13. $x \lll y$ and $|u| < |x| + |v|$ implies $|u + x| < |y| + |v|$,

14. $x \ll y$ and $0 < y$ implies $0 < x + y$,
15. $x \ll y$ and $x \ll z$ and $r(y + z) \neq 0$ implies $x \ll r(y + z)$,
16. $x \ll y$ and $x \ll z$ and $y + z \neq 0$ implies $x \ll y + z$,
17. $x \ll z$ or $y \ll z$ implies $r(x + y) \ll z$,
18. $y \lll z$ and $z \lll x$ and $x > 0$ and $|u| < |y| + |z|$ implies $(u + y) + x > 0$,
19. $u \lll v \lll x$ implies $u \lll w$ or $w \lll x$.

Axioms 15–17 summarize Lemma 6.1 of Kulisch and Miranker [2, p. 293]. There are more axioms, but they cannot be conveniently stated without further notation.

Now let $(S, \leqslant, +, -)$ be a model for these axioms (including the several axioms not yet stated). Our main theorem is that S can be extended to an ordered group with a rounding function onto S. Let us begin the reconstruction of this group. From the standard model we will recover the dyadic rationals, i.e., those rationals whose denominator is a power of two. By a freak of nature, this does not turn out to be a field. In general, we must consider the set of all finite sequences from S factored by the equivalence relation of having the same exact sum in the extended group. Our problem is to define this relation without reference to any structure external to S.

A sequence $\langle x_0, \ldots, x_n \rangle$ of elements from S is in normal form if, for all $i < n$, we have $x_{i+1} \lll x_i$, and it is in weak normal form if for all $i < n$ either x_{i+1} is zero or $x_{i+1} \ll x_i$. Weak normal form is an auxiliary concept used solely on the road to normal form. In the standard model, a normal form is completely determined by its exact sum, since $x_0 = O(s)$, $x_1 = O(s - x_0)$, etc. Now consider an arbitrary sequence $\langle x_0, \ldots, x_n \rangle$. If we replace the pair x_i, x_j by the new pair $x_i + x_j$, $r(x_i + x_j)$, then the exact sum is unchanged. The principal theorem of this research is that such reductions can always be used to reduce an arbitrary sequence to normal form, and that the normal form achieved is independent of the particular sequence of reductions used to derive it. It is even independent of the order of the original sequence. This result will follow solely on the basis of explicit first order axioms for S and will not require any external structure. However, by the external considerations just discussed, we already know it to be true in the standard model. Existence will follow from a modification of the Bohlander algorithm presented by Kulisch and Miranker.

Let us denote the single reduction on the pair x_i, x_j by R_{ij}. Using a right-hand function notation, this means

$$\langle \cdots x_i \cdots x_j \cdots \rangle R_{ij} = \langle \cdots x_i + x_j \cdots r(x_i + x_j) \cdots \rangle.$$

We then use the notation RS to mean first do R then do S. A derivation D is a finite product of reductions. We must prove two things: for every finite sequence s there is a derivation D such that sD is a normal form, and further that if sD and sD' are both normal forms then $sD = sD'$.

Let us begin with three element sequences. Let $E_k = R_{k-1,k} * R_{k-2,k-1} * \cdots * R_{12} * R_{01}$ and let $D_n = E_1 * \cdots * E_n$. We claim that for all x, y, z, $\langle x, y, z \rangle D_2$ is a weak normal form. To see this let us draw some pictures. We employ the picture $x \leftarrow y$ to mean that either $x = 0$ or $y \ll x$. This picture is to be interpreted transitively even though the relation it represents is not transitive, i.e., $x \leftarrow y \leftarrow z$

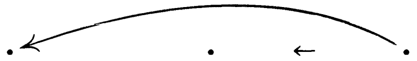
implicitly asserts $x \leftarrow z$. We start with three unrelated points.

• • •

Do R_{01} and read axiom 11. Get

• ← • •

Then R_{12} with axioms 11 and 17 yield,



Then R_{01} again and axioms 11, 15, and 16 yield,

• ← • ← •

Let us now investigate the process by which three element weak normal forms may be converted into normal forms. In the first case, it is obvious that if any of the three elements are zero, another pass with E_2 will produce a normal form. So suppose $x \ggg y \gg z$ is the weak normal form produced by the above process and that x, y, z are all nonzero. If $x \ggg y + z$, then one application of R_{12} produces normal form. The only way that $x \ggg y + z$ can fail is that y is exactly halfway between x and its immediate neighbor and that z has the same sign as y. In that case, after doing R_{12}, an application of R_{01} changes x to its neighbor and changes the sign of $y + z$ while decreasing its absolute value. In other words an application of E_2 leaves nearly the same situation as we started with except that y is no longer exactly halfway between x and its neighbor. Thus another application of R_{12} produces normal form. We can summarize this with an axiom:

$$20. \quad D_2E_2R_{12}R_{01} = D_2E_2R_{12}.$$

By the above discussion, this axiom implies that if $\langle x, y, z\rangle$ is a weak normal form then $\langle x, y, z\rangle E_2R_{12}$ is a normal form.

In the standard model, the normal form is unique. This implies that $R_{12}D_2E_2R_{12} = D_2E_2R_{12}$ since s and sR_{12} have the same normal form. We also know that if $x_j \lll x_i$ then R_{ij} has no effect on the sequence. This follows directly from the definition of r given by Knuth. We have another axiom:

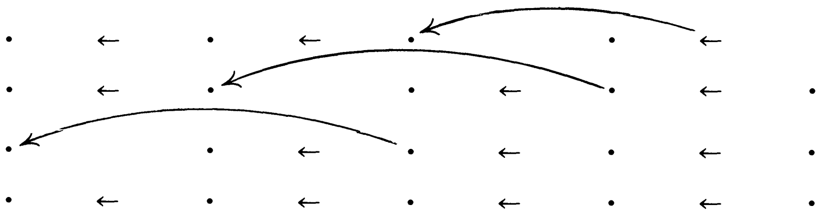
$$21. \quad R_{12}D_2E_2R_{12} = D_2E_2R_{12}.$$

This implies the uniqueness of normal form for three element sequences. For if sD is a normal form, then $sDR = sD$. Thus $sDD_2 = sD$. On the other hand, axiom 21 implies that $sDD_2 = sD_2$, so we see that sD_2 is the only normal form for s.

We now turn to the general case, $s = \langle x_0, \ldots, x_n\rangle$. The outline of this case is exactly the same as the three element case. To prove the first step that sD_n is a weak normal form, it clearly suffices to prove that if $\langle x_0, \ldots, x_n\rangle$ is a weak normal form, then for any x_{n+1}, $\langle x_0, \ldots, x_n, x_{n+1}\rangle E_{n+1}$ is also a weak normal form. This easily follows from a Pacman argument. We start with the picture,

• ← • ← • ← • •

This picture will be successively transformed through the following sequence by E_{n+1}:



(See axioms 11, 15, 16, 17.)

Let us now examine the process by which weak normal form is converted into normal form. Define D'_n similarly to D_n except that in D'_n each E_k is done twice, i.e., $D'_n = E_1^2 E_2^2 \cdots E_n^2$. The next lemma says that if s is a weak normal form of length n, then sD'_n is a normal form.

LEMMA. *If $\langle x_0, \ldots, x_n \rangle$ is a normal form and $x_{n+1} \ll x_n$, then $\langle x_0, \ldots, x_n, x_{n+1} \rangle E_{n+1} E_{n+1}$ is a normal form.*

Proof. Let us set $x_{n+1} = u$, and $x_n = x$, and $x_{n-1} = y$, and $x_{n-2} = z$. If $u + x \lll y$ or even if $(u + x) + y \lll z$, the lemma easily follows from the above discussion of three element sequences. In the standard model, the remaining case is somewhat special. We must have that u, x, y all have the same sign and u is a power of two and y is all ones and x, y, z are nestled together as close as possible. In that case we can prove the following two axioms by simple arithmetic.

22. $u \ll x \lll y \lll z$ implies $(u + x) + y \ll z$,

23. $u \ll x \lll y \lll z$ implies $r((u + x) + y) \lll r(((u + x) + y) + z)$.

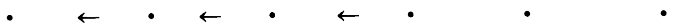
These two axioms clearly say that the lemma now follows by induction from the case $n = n - 2$.

There is now a standard method for reducing any sequence to normal form. If s is a sequence of length $n + 1$, then $sD_nD'_n$ is a normal form. We must now show that this normal form is unique, i.e., that any other sequence of reductions leading to a normal form must produce exactly the same result. This requires several lemmas. Let us define two weak normal forms, s and t, to be equivalent if there is a sequence of weak normal forms $s_0, \ldots, s_n$ with $s_0 = s$ and $s_n = t$ and having the property that for $i < n$, s_i and s_{i+1} differ at only two consecutive places and at those places they have the same sum and same residue, i.e., for $i < n$ there is a j such that $s_i R_{j,j+1} = s_{i+1} R_{j,j+1}$. We first show that if s is a one-step reduction of t (i.e. $s = tR_{i,i+1}$), then sD_n and tD_n are equivalent weak normal forms. Since this proposition is already known to be true in the standard model, we can postulate its truth for sequences of length less than five.

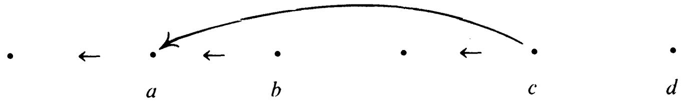
24. $R_{34} D_4 D'_4 = D_4 D'_4$.

LEMMA. *If $x + y = x' + y'$ and $r(x + y) = r(x' + y')$ and $\langle x_0, \ldots, x_n \rangle$ is a weak normal form, then $\langle x_0, \ldots, x_n, x, y \rangle E_{n+1} E_{n+2}$ and $\langle x_0, \ldots, x_n, x', y' \rangle E_{n+1} E_{n+2}$ are equivalent weak normal forms.*

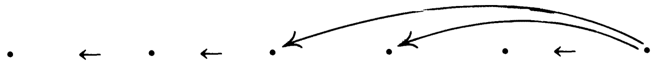
Proof. The pictoral representation of $\langle x_0,\ldots, x_n, x, y\rangle$ is



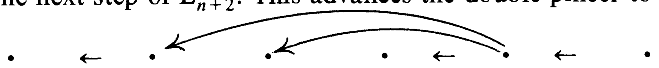
The first step of E_{n+1} yields



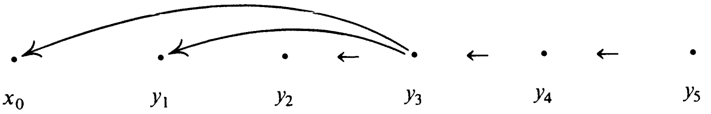
At this point we should add a to b, continuing the application of E_{n+1}, but E_{n+1} will never again alter the value of c. Thus we may as well do the first step of E_{n+2} before proceeding with E_{n+1}. This gives



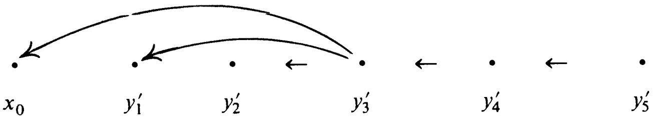
We must now continue with the next step of E_{n+1}, but again this can be immediately followed by the next step of E_{n+2}. This advances the double pincer to



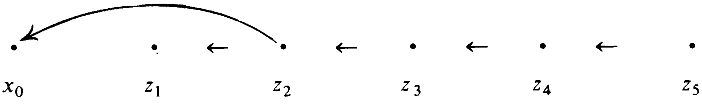
Continuing, we see that $\langle x_0,\ldots, x_n, x, y\rangle$ can be reduced to



where the value x_0 has not yet been altered. Similarly, $\langle x_0,\ldots, x_n, x', y'\rangle$ can be advanced to



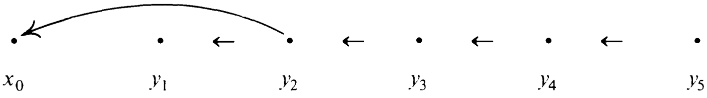
By the inductive hypothesis, we may assume that $\langle y_1+y_2, r(y_1+y_2), y_3,\ldots\rangle$ is equivalent to $\langle y_1'+y_2', r(y_1'+y_2'), y_3',\ldots\rangle$. In both cases we are supposed to proceed by applying $R_{01}R_{12}R_{01}$. However by axiom 24, we get equivalent results using $R_{12}R_{01}R_{12}R_{01}$ instead. First doing R_{12} has the effect of producing



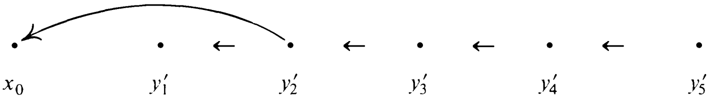
and similarly for $\langle x_0, z_1', z_2',\ldots\rangle$, where our inductive hypothesis says that $\langle z_1,\ldots, z_{n+2}\rangle$ and $\langle z_1',\ldots, z_{n+2}'\rangle$ are equivalent weak normal forms. Both of these sequences are to be reduced with $R_{01}R_{12}R_{01}$. We must show that the results are equivalent. We may as well assume that $\langle z_1,\ldots, z_{n+2}\rangle$ and $\langle z_1',\ldots, z_{n+2}'\rangle$ are immediately equivalent, that is to say, we may assume that they differ at only two consecutive places, etc. If $z_1+z_2=z_1'+z_2'$ and $r(z_1+z_2)=r(z_1'+z_2')$, then axiom 24 again implies that the results are equivalent. Thus we may as well assume that $z_k=z_k'$ for $k=1,4,5,\ldots$ and that $z_2+z_3=z_2'+z_3'$ and $r(z_2+z_3)=r(z_2'+z_3')$. Then axiom 24 again says the results are equivalent.

Lemma. *If $s = \langle x_0, \ldots, x_n \rangle$ and $t = \langle x_0', \ldots, x_n' \rangle$ are equivalent weak normal forms, then $\langle s, x \rangle E_{n+1}$ and $\langle t, x \rangle E_{n+1}$ are also equivalent.*

Proof. As usual,we may as well assume that s and t differ at only two consecutive places, etc. We may also assume that $x_n \neq x_n'$ or $x_{n-1} \neq x_{n-1}'$, for otherwise the inductive hypothesis would be immediately applicable. Therefore $x_0 = x_0'$. Now in both cases perform all but the last step of E_{n+1} to get



and



By the inductive hypothesis, $\langle y_1, \ldots, y_{n+1} \rangle$ and $\langle y_1', \ldots, y_{n+1}' \rangle$ are equivalent weak normal forms. Now we proceed just as in the last lemma. Namely, again we may assume that these two sequences differ at only two consecutive places, etc., and then apply axiom 24.

The next step in our argument is to prove that two equivalent weak normal forms have the same normal form. This is true in the standard model, so we can use any special cases as axioms, i.e.

25. If s is a weak normal form of length 5, then $sD_4' = sR_{34}D_4'$.

Lemma. *If s and s' are equivalent weak normal forms of length $n + 1$, then $sD_n' = s'D_n'$.*

Proof. The proof is by induction on n. In the first place, we may as well assume that s and s' differ at only two consecutive places and at those two places they have the same sum and same residue. In fact, we may as well assume that they differ at only the last two places and prove that $sE_{n-1}^2E_n^2 = s'E_{n-1}^2E_n^2$. Imagine the partial completion of these reductions on s. First do all but the very last step (R_{01}) of E_{n-1}. Then since R_{01} only changes x_0 and x_1, we may begin the application of the second E_{n-1}, doing all but the last two steps ($R_{12}R_{01}$). Now do all but the last three steps of the first E_n and all but the last four steps of the final E_n. Do the same thing to s'. We now have two sequences, t and t', with the property that

$$sE_{n-1}^2E_n^2 = tR_{01}R_{12}R_{01}R_{23}R_{12}R_{01}R_{34}R_{23}R_{12}R_{01}$$

and similarly for s' and t'. By the inductive hypothesis and our assumption that s and s' have the same first element, if we leave out the R_{01} reductions, t and t' will give equal results. Therefore the sequences t and t' are equal from coordinate 5 on and the first five terms are equivalent. From axiom 25, we already know that normal form is unique for sequences of length 5. Thus the lemma is proven.

Theorem. *Every finite sequence from S can be reduced to normal form with reductions of the form $R_{i,j}$ and the normal form achieved is independent of the particular sequence of reductions used as well as the order of the original sequence.*

Proof. The last three lemmas easily imply that normal form can be achieved with reductions of the form $R_{i,i+1}$ and that provided just these reductions are used, the normal form is unique. This in turn implies that the normal form is independent of the order of the original sequence since the reduction $R_{i,i+1}$ obliterates the transposition $(i, i+1)$ and every permutation is a product of such transpositions. As a consequence of this we can also remove the restriction $j = i + 1$.

Having completed the proof of the normal form theorem, we can now define the reconstructed group.

Definition. For s and t finite sequences from S,

$s + + t = s$ concatenate t,
$O(s) =$ the dominant term of the normal form,
s is positive iff $O(s) > 0$.

The normal form theorem implies that $++$ is well defined with respect to the equivalence relation of having the same normal form. Our next task is to verify the axioms of ordered groups. The group axioms follow immediately from the definitions and the normal form theorem and can be left to the reader. What we have left to show is that the sum of positives is positive and that O is order preserving.

LEMMA. *If $\langle y, y_1, \ldots, y_n \rangle$ and $\langle y, x_1, \ldots, x_m \rangle$ are both normal forms and $y > 0$, then $\langle y, y_1, \ldots, y_n, x_1, \ldots, x_m \rangle$ is positive.*

Proof. The proof is by induction on m. Let $s = \langle y, y_1, \ldots, y_n, x_1 \rangle E_{n+1}$. Then s is a weak normal form and axioms 13 and 18 say it has a positive leading term. In order to apply the inductive hypothesis, we need to know that $x_1 \lll O(s)$. The problem is that even though axioms 13 and 18 give a good handle on the leading term of s, they do not say as much about $O(s)$ because s is merely a weak normal form, not a normal form. We do know that the second term of s is $\lll$ the leading term. What is needed, therefore, is an axiom saying that the leading term of such a weak normal form does not change much on reduction to normal form. But how can it change at all? The axiom $x \ll y \lll z$ implies $x + y \lll z$ is almost true. The only exception occurs when x and y have the same sign and y is a power of two nestled as close as possible to z. In that case $(x + y) + z$ is either the immediate predecessor or immediate successor of z depending on the sign of y. We can entirely summarize the state of affairs with the following axioms:

26. Every nonzero element has an immediate predecessor and an immediate successor, but zero has neither.
27. $x \ll y \lll z$ and $y < 0$ implies $(x + y) + z = z$ or $(x + y) + z$ is the predecessor of z.
28. $x \ll y \lll z$ and $y > 0$ implies $(x + y) + z = z$ or $(x + y) + z$ is the successor of z.
29. $x \ll y \lll z$ implies $(x + y) + z = z$ or $r((x + y) + z)$ and y have opposite signs.

With these axioms it is easily proven that if s is a weak normal form with second term $\lll$ its leading term, then $O(s)$ is at worst the successor or predecessor of its leading term. Thus the fact that we want $x_1 \lll O(s)$ can now be easily formulated as an axiom.

30. $y_1 \lll y_0$ and $x_0 \lll y_0$ and $0 < y_0$ and $|u| < |y_1| + |x_0|$ and
$x_1 \lll x_0$ implies $x_1 \lll$ both the successor and predecessor of $(u + y_1) + y_0$.

Our goal is to prove two things: that the sum of positive elements is positive and that O is order preserving. Both these goals are accomplished by the next lemma.

LEMMA. *If* $0 < O(s) + O(t)$, *then* $s + + t$ *is positive.*

Proof. Let $s = \langle x_0, \ldots, x_n \rangle$ and $t = \langle y_0, \ldots, y_m \rangle$ be normal forms. If $x_0 \lll y_0$, then the previous lemma is applicable and the present lemma is proven. Therefore, by axiom 19, we may assume that $y_2 \lll x_0$. What we propose to do is to let $t' = s + + \langle x_0 \rangle$ and $s' = \langle x_1, \ldots, x_n \rangle$ and prove $O(s') + O(t')$ is also positive. The lemma would then follow by induction on n. To this end, let us first decompose t into $\langle y_0, y_1 \rangle + + \langle y_2, \ldots, y_m \rangle$. Let $\langle z_0, z_1, z_2 \rangle$ be the normal form for $\langle y_0, y_1, x_0 \rangle$. In the standard model, we know that $x_1 + z_0$ is positive. The problem with just formulating this as a new axiom is that $\langle z_0, z_1, z_2, y_2, \ldots, y_n \rangle$ is not necessarily a normal form. Quite possibly the value z_0 may be changed during the reduction to normal form. We must formulate an axiom which guarantees that this changed value still has $z_0 + x_1$ positive. Let us note however that since $y_2 \lll x_0$ the above sequence, $\langle z_0, z_1, z_2, y_2, \ldots, y_n \rangle$, is a weak normal form. Most of the time z_0 will have a nonzero digit at least as significant as the least significant digit of x_0. In that case, no matter how the reduction to normal form goes, it cannot change the sign of $z_0 + x_1$. The only border line case is when x_0 is a power of 2, and y_0 is the negative of its successor, and y_1 is a power of 2 nestled as close as possible to y_1 with the same sign as x_0. In that case, $(x_0 + y_1) + y_0 = 0$, so that $z_0 = -y_1$, $z_1 = 0$, and $z_2 = 0$. But then $y_2 \lll z_0$ and z_0 does not change at all. Based on this discussion, we can easily see that the following axiom proves the lemma:

31. If $x_1 \lll x_0$ and $y_1 \lll y_0$ and $0 < x_0 + y_0$ and $z = O(y_0, y_1, x_0)$,
then either $(x_0 + y_1) + y_0 = 0$ and $r((x_0 + y_1) + y_0) = -y_1$ or for
any $u \ll z$ and any neighbor z' of $u + z$, we have $x_1 + z'$ is positive.

THEOREM. *If* $\langle S, \leqslant, +, - \rangle$ *satisfies axioms* 1–31, *then there is an ordered group* G *extending* S *and a rounding function* O *defining* S.

We now turn to the problem of making the extended group into a field. Our plan is to first introduce axioms for multiplication which allow us to add a multiplication operation to G obtaining an integral domain. We will then add axioms for division which will allow us to extend the function O to the fraction field. We begin with the obvious axioms.

32. $xy = yz$,
33. $x \leqslant y$ and $0 \leqslant z$ implies $xz \leqslant yz$,
34. $xy = 0$ implies $x = 0$ or $y = 0$,
35. $x1 = x$: $x0 = 0$,
36. $-(xy) = (-x)y = x(-y)$,
37. $(x + x) - x = x$,
38. $x \lll y$ implies $8|xz| < |yz|$.

In axiom 38, we use the notation $2x = x + x$, $4x = 2x + 2x$, $8x = 4x + 4x$. Axiom 37 says that there is no error in these calculations, i.e., $r(x + x) = 0$.

We need a residue function r_* for multiplication. Since the product of single precision numbers is exactly representable in double precision, there is no problem with the existence of this function. However I do not as yet know of a definition of r_* similar to Knuth's definition of r_+. We must therefore use r_* as a new primitive. In the standard model, it is defined via the equation,

$$r_*(x, y) + + xy = x * * y.$$

We use the notation $r(xy)$ for $r_*(x, y)$. This leads to anomalies such as $r(x(y + z))$ for $r_*(x, y + z)$, but it seems better than the alternatives. Just like addition, our main task is to axiomatize this function. Here are some more trivial axioms:

39. $r(xy) \lll xy$,
40. $r(x1) = r(x0) = 0$,
41. $r((-x)y) = r(x(-y)) = -r(xy)$,
42. $8|r(xy)| < |xy|$.

The definition of $s * * t$ forces itself upon us.

Definition.

$s * * t =$ a sequence of all terms of the form xy or $r(xy)$ such that x is a term from s and y a term from t.

We need not specify order in this definition since the normal form theorem implies that it is irrelevant. The next axiom says that this operation is well defined with respect to the equivalence relation of having the same normal form.

43. $\langle xz, r(xz), yz, r(yz)\rangle D_3D_3' =$
$\langle (x + y)z, r((x + y)z), r(x + y)z, r(r(x + y)z)\rangle D_3D_3'$.

This axiom says that doing a reduction before a multiply does not change the equivalence class of the result. The distributive law follows immediately from the definitions, but the associative law needs a new axiom.

44. $\langle (xy)z, r((xy)z), r(xy)z, r(r(xy)z)\rangle D_3D_3' =$
$\langle x(yz), r(x(yz)), xr(yz), r(xr(yz))\rangle D_3D_3'$.

All that is left to complete the integral domain verification is to show that the product of positives is positive. This will incidently establish that $st = 0$ implies $s = 0$ or $t = 0$.

LEMMA. *If $s = \langle x_0, \ldots, x_n\rangle$ is a normal form, then $|\langle y\rangle s| \leqslant 2|yx_0|$.*

Proof. $\langle y\rangle s = \langle y\rangle\langle x_0\rangle + + \langle y\rangle\langle x_1, \ldots, x_n\rangle = \langle yx_0\rangle + + \langle r(yx_0)\rangle + + \langle y\rangle\langle x_1, \ldots, x_n\rangle$. Therefore, we may take absolute values and apply axioms 37, 38, 42, and the inductive hypothesis to get

$$|\langle y\rangle s| \leqslant |\langle yx_0\rangle| + + |\langle yx_0\rangle| = 2|\langle yx_0\rangle|.$$

In the future, we will not even attempt to put $\langle\ \rangle$ around singletons. The reader will have to decide from context whether or not an element from S is being used as a singleton sequence in G.

LEMMA. *If $s = \langle x_0, \ldots, x_n\rangle$ and $t = \langle y_0, \ldots, y_n\rangle$ are normal forms, then $|st| \leqslant 4|x_0 y_0|$.*

Proof. As in the previous lemma.

LEMMA. *If s and t are positive, then so is st.*

Proof. This proof actually involves the summation of a geometric series. Let $s = \langle x_0, \ldots, x_n\rangle$ and $t = \langle y_0, \ldots, y_n\rangle$ be normal forms with $x_0, y_0 > 0$. Then

$$\begin{aligned} st = {} & \langle x_0\rangle\langle y_0\rangle + + \langle x_0\rangle\langle y_1, \ldots, y_n\rangle \\ & + + \langle y_0\rangle\langle x_1, \ldots, x_n\rangle + + \langle x_1, \ldots, x_n\rangle\langle y_1, \ldots, y_n\rangle. \end{aligned}$$

Thus st is at least as large as

$$x_0 y_0 - |r(x_0 y_0)| - 2|x_0 y_1| - 2|x_1 y_0| - 4|x_1 y_1|.$$

Let $z = \max(|r(x_0 y_0)|, 2|x_0 y_1|, 2|x_1 y_0|, 4|x_1 y_1|)$. By axioms 37, 38, 39, $4z < |x_0 y_0|$, but st is at least as large as $|x_0 y_0| - 4z$. This proves the lemma and completes the proof of the following theorem:

THEOREM. *If the algebra $\langle S, \leqslant, 0, 1, +, -, *, r_*\rangle$ satisfies axioms 1–44, then there is an ordered integral domain extending S and a rounding function O defining S.*

Our final step is to extend O to the fraction field of this integral domain. We shall do this not by adding division as a primitive to S, but rather by adding enough axioms so that division can be defined from multiplication. A preliminary definition of x/y would be the largest z such that $zy \leqslant x$. If the rounding were truncation, this would be the final definition as well. We must give first order axioms to guarantee the existence of such a largest element. The obvious approach of postulating that every bounded set has a least upper bound has the disadvantage that it is not a first order axiom and so would require a whole new set of axioms for set existence, thus opening us to the supplications of various snake oil salesmen with their bottles of measurable cardinals, etc.

Axiom 26 in itself does not guarantee the existence of the largest z such that $zy \leqslant x$. We must somehow postulate the existence of a finite interval (z_1, z_2) such that $z_1 y \leqslant x$ and $z_2 y > x$. Let x and y be arbitrary positive elements from S. Let $x + dx$ be the successor of x and let $y - dy$ be the predecessor of y. In the standard model we seek an upper bound for the number of screen points in the interval from x/y to $(x + dx)/(y - dy)$. Let $z = x/y$ and let $dz = (x + dx)/(y - dy) - x/y$. From Taylor's theorem,

$$\begin{aligned} |dz/z| \leqslant {} & |dx/x + dy/y| \\ & + |dx/x + dy/y|\,\left|y^2/(y - dy)^2\right|\,|dy/(y - dy)|. \end{aligned}$$

We also know that $|dx/x|$ is between 2^{-24} and 2^{-23}. The same for dy/y and $dy/(y - dy)$. Thus

$$|dz/z| \leqslant 2^{-22}(1 + 2^{-22}).$$

Thus we see that at most five screen points can fit in between z and $z + dz$. We can formulate an axiom summarizing the foregoing.

45. If x and y are both positive and $x + dx$ is the successor of x and $y - dy$ is the predecessor of y, then there exist z_1 and z_2 such that $z_1 y \leqslant x$ and $z_2(y - dy) > x + dx$ and there are at most five points strictly between z_1 and z_2.

LEMMA. *For any two positive sequences s and t from G, there is a largest screen point z such that $\langle z \rangle t \leqslant s$.*

Proof. Choose x and y so that x and $x + dx$ straddle s and likewise for y, $y - dy$, and t.

Definition.

$$\mathrm{Tr}(s, t) = \text{the largest } z \text{ such that } \langle z \rangle t \leqslant s,$$

$$\mathrm{Tr}^+(s, t) = \text{the successor of } \mathrm{Tr}(s, t).$$

The function Tr is of course the truncation of s/t. Obviously, $st' \leqslant s't$ implies $\mathrm{Tr}(s, t) \leqslant \mathrm{Tr}(s', t')$, i.e.,Tr is defined and order preserving on the fraction field, but it does not extend O.

Definition. $O(s, t) = \mathrm{Tr}^+(s, t)$ if there is a z such that $z \lll \mathrm{Tr}^+(s, t)$ and $\langle Tr^+(s, t), z \rangle t \leqslant s$, and $O(s, t) = \mathrm{Tr}(s, t)$ otherwise.

LEMMA. *If $st' \leqslant s't$, then $O(s, t) \leqslant O(s', t')$.*

Proof. In view of the fact that Tr and Tr^+ both satisfy this condition, it suffices to consider only the case when $\mathrm{Tr}(s, t) = \mathrm{Tr}(s', t')$. This case follows easily from the definition.

As a corollary to this lemma we may conclude that O is defined on the fraction field and is order preserving. All that is left to show is that $O(s, 1) = O(s)$. But again this is an immediate consequence of the definition. Thus we have concluded the proof that has occupied the last 20 pages.

THEOREM. *If $\langle S, \leqslant, 0, 1, +, -, *, r_* \rangle$ satisfies axioms 1–45, then there is an ordered field F extending S and a rounding function O mapping F onto S satisfying the conditions given on the second page of this paper.*

Department of Mathematics
Pennsylvania State University
University Park, Pennsylvania 16802

1. DONALD E. KNUTH, *The Art of Computer Programming*, Vol. 2, Addison-Wesley, Menlo Park, Calif., 1969.

2. U. W. KULISCH & W. L. MIRANKER, *Computer Arithmetic in Theory and Practice*, Academic Press, New York, 1981.

3. L. B. RALL, *Accurate Arithmetic for Scientific Computation*, Proceedings of the 1982 Army Numerical Analysis and Computer Conference, 1982.

4. J. H. WILKENSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N. J., 1963.