# A Formalization of Floating-Point Numeric Base Conversion

## DAVID W. MATULA

*Abstract*—The process of converting arbitrary real numbers into floating-point format is formalized as a mapping of the reals into a specified subset of real numbers. The structure of this subset, the set of $n$ significant digit base $\beta$ floating-point numbers, is analyzed and properties of conversion mappings are determined. For a restricted conversion mapping of the $n$ significant digit base $\beta$ numbers to the $m$ significant-digit base $\delta$ numbers, the one-to-one, onto, and order-preserving properties of the mapping are summarized. Multiple conversions consisting of a composition of individual conversion mappings are investigated and some results of the invariant points of such compound conversions are presented. The hardware and software implications of these results with regard to establishing goals and standards for floating-point formats and conversion procedures are considered.

*Index Terms*—Accumulated error of successive conversions, base conversion, "equivalent digits" formula, floating-point numbers, mixed-base computational environments, round-off error, significant digits.

## I. INTRODUCTION AND SUMMARY

THE necessity of base conversion of numeric data during some stages of computation on a digital computer is a de facto component of practical numeric computation that must be recognized in any complete analysis of digital computation. On the hardware level, the trend towards establishment of computer networks with possibly differently based machines and, on the software level, the mixed-base flexibility inherent in the PL/I language specifications, both suggest that internal data of certain jobs may be necessarily subjected to multiple conversions before job termination. Thus, references to a purportedly constant floating-point datum occurring at different points during program execution might encounter altered datum values. Both hardware and software designers must recognize this problem, and each can benefit from the fundamental principles obtained by considering base conversion as a mathematical transformation. In this report we shall follow the notation of our previous work [1]–[4], integrating the mainstream of results from those articles with a general formal development of base conversion, and providing new results particularly in the area of multiple conversions.

A fundamental analysis of base conversion is concerned

first with determination of the theoretical limitations inherent in any implementation of base conversion. The actual algorithmic mechanics of any theoretically realizable conversion procedure is then a secondary (albeit nontrivial [5]) problem which will not concern us in this paper. The formalization we introduce provides the vehicle for studying the properties and recognizing the inherent anomalies of the conversion process, which must necessarily then guide the performance specifications for floating-point representations and base conversions at both hardware and software levels.

Converting integer and fixed-point data to an "equivalent" differently based number system is generally achieved by utilizing essentially $\log_\delta \beta$ times as many digits in the new base $\delta$ as were present for representing numbers in the old base $\beta$ system. This simplified notion of equivalence does not extend to the conversion of floating-point systems. Actually, conversion between floating-point number systems introduces subtle difficulties peculiar to the structure of these systems so that no such convenient formula for equating the "numbers of significant digits" is even meaningful. Thus, our formalization of floating-point base conversion is preceded in Section II by a careful analysis of floating-point number systems.

Following our previous work [1]–[4], a system of floating-point numbers of $n$ significant digits to the base $\beta$ is characterized as a *significance space*, $S_\beta^n$, and in Theorem 1 the number of elements in $S_\delta^m$ relative to the number of elements in $S_\beta^n$ within a specified interval is shown to converge to $((\delta-1)\delta^{m-1}/((\beta-1)\beta^{n-1})) \log_\delta \beta$ as the interval grows to include the whole real line. This relative density of "total membership" of two significance spaces provides only a gross comparison of the two number systems, for actually there is considerable local variation in the relative density. The *gap function* [1] $\Gamma_\beta^n(x)$ is defined as the relative difference between nearest neighbors of $S_\beta^n$ at $x$. A comparison of the graphs of $\Gamma_\beta^n$ and $\Gamma_\delta^m$ provides more insight into the comparability of two differently based floating-point systems than any simplified "equivalent digit" formula.

Having characterized floating-point number systems and the gap function, the conversion of the real numbers into $S_\beta^n$ both by rounding and by truncation procedures are then formalized in Section III. The order-preserving properties of these conversion mappings are detailed, and the base conversion theorem [2] is stated, which gives the necessary

and sufficient conditions for a conversion mapping from $S_\beta^n$ to $S_\delta^m$ to be 1) one-to-one and 2) onto.

The important problems associated with multiple conversions of a datum are analyzed in Section IV. The composition of repeated rounding and/or truncation conversions is termed a *compound conversion*, and the associated *invariant points* (i.e., the points mapped into themselves) of such a compound conversion mapping are analyzed. For a compound truncation conversion, it is shown that the only invariant points of the mapping are the numbers common to all of the significance spaces involved. Thus, considerable importance must be attached to the intersection of significance spaces, and in Theorems 5 and 6 these intersections are shown to exhibit a crucial dependence on the commensurability of the bases. Specifically, significance spaces with commensurable bases will always jointly contain a common significance space. For example, the six-digit hexadecimal numbers and the eight-digit octal numbers both contain all 21-bit binary numbers. On the other hand, the members common to significance spaces with incommensurable bases (e.g., binary and decimal) will be finite in number and, for cases of computational interest, these members will typically all fall in an interval much smaller than the interval range provided by current exponent ranges on digital computers.

As a consequence of the limited membership of points common to two incommensurable significance spaces, the multiple back-and-forth conversion of a "constant datum" between two incommensurable significance spaces by truncation conversion can accumulate error so as to invalidate even the leading digit of the value of this "constant datum." Practically, the process of updating a binary coded decimal (BCD) tape on a binary machine might well subject stored data which is never updated to multiple binary-decimal conversions, so BCD tape updating is very sensitive to such anomalies of compound conversions. Fortunately, we can show that under very general conditions, iterated rounding conversion of a datum between two significance spaces quickly generates a stable pair of values, each of which is a reasonable approximation of the initial datum. Under the stronger conditions given in the in-and-out conversion theorem [3], rounding conversion through an intermediate significance space can be guaranteed to regenerate the initial datum of the original significance space.

In the presence of more than two incommensurable bases, the possibility of cyclic conversions' accumulating error in a datum is shown to exist even under rounding conversion. Our final result resolves the problem of controlling the overall growth of accumulated conversion error within a mixed-base computational environment by a process which standardizes a datum's value in each of the significance spaces involved.

## II. Floating-Point Number Systems

A formalization of floating-point number systems must start with a characterization of the set of floating-point numbers, preferably divorced from the cumbersome digit sequence representational notation. In previous article [1]–[4] this set has been termed a significance space.

*Definition:* For the integers $\beta \geq 2$, called the *base*, and $n \geq 1$ called the *significance* (or *precision*), let the *significance space* $S_\beta^n$ be the following set of real numbers:

$$S_\beta^n = \{b | b = k\beta^j \text{ for some integers } k, j \text{ where } |k| < \beta^n\}.$$

For clarity we shall utilize the Greek letters $\alpha$, $\beta$, and $\delta$ to denote bases; the English letters $a$, $b$, $c$, and $d$ will denote elements of a significance space, i.e., the so called "floating-point" numbers; the letters, $i, j, k, l, m, n, p$, and $q$ will denote integers; and $x$, $y$, and $z$ will denote arbitrary real numbers.

For an element $b = k\beta^j \in S_\beta^n$ the actual floating-point representation of $b$ can be visualized as having the fixed-point integer portion $k$ represented by a sign and $n$ or less digits to the assumed base $\beta$, with the exponent portion then represented by the integer $j$. The floating-point representation just described will not in general be unique, with digit sequence realizations of $b$ corresponding to both "normalized" and "unnormalized" forms possible in some cases. Considerations related to the nonunique determination of $k, j$ in $b = k\beta^j$ will be treated where necessary; however, our main concern is with membership of $b$ in the set of real numbers $S_\beta^n$ for which the form of representing $b$ is irrelevant. Note that the significance space $S_\beta^n$ differs from an actual floating-point number system in that there is no bound on the exponent portion $j$ of the members $b = k\beta^j \in S_\beta^n$. Thus $S$ is actually an infinite set. Since we shall not concern ourselves with underflow and overflow problems, the significance space $S_\beta^n$ is a perfectly acceptable model of a floating-point number system for our purposes.

It is easy to visualize the change in the set $S_\beta^n$ caused by varying significance, since increasing $n$ to $n+1$ maintains all members of $S_\beta^n$ and adds $\beta - 1$ new members uniformly spaced between every neighboring pair of members of $S$. The dependence of the membership of $S_\beta^n$ on the base $\beta$ is far more subtle. In practice it has been convenient to identify a nondecimal floating-point number system with the "appropriate" decimal based system; however, we now show that such a purported equivalence glosses over certain inherent anomalies between differently based floating-point number systems.

A gross comparison of two differently based significance spaces can be obtained by determining the relative number of members of each space over a comparable range. For example, the 3 (significant)-bit binary numbers between unity and one thousand are 1, $1.01_2 = 1.25$, $1.10_2 = 1.5$, $1.11_2 = 1.75$, $10.0_2 = 2$, $10.1_2 = 2.5$, $11.0_2 = 3$, $11.1_2 = 3.5$, $100_2 = 4$, $101_2 = 5, \cdots$, $1100000000_2 = 768$, $1110000000 = 896$; and the 1 (significant)-digit decimal numbers over the same range are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000. In Fig. 1 these members are indicated by tickmarks on the real line plotted on a logarithmic scale so that the log periodic nature of the spacing between floating-point numbers is evident. The ratio of the numbers of members of these systems over this interval is $40/28 = 1.43$. Now this ratio
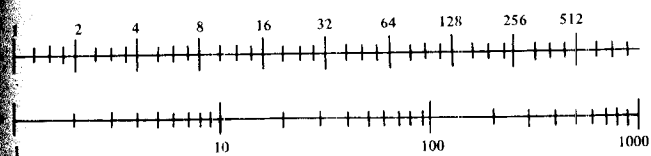
Fig. 1. Tickmarks plotted on a log scale showing (a) the 40 3 (significant)-bit binary numbers and (b) the 28 1 (significant)-digit decimal numbers over the range $[1, 10^3]$.

membership density will vary with the choice of interval; however, a reasonable overall comparison of any two floating-point systems can be calculated by determining the limit of such a ratio as the interval grows to include the whole real line.

*Theorem 1:* Let $S_\beta^n$ and $S_\delta^m$ be any two significance spaces. Then letting $|S|$ denote the number of members of the set $S$,

$$\lim_{M \to \infty} \left| \frac{\left\{ d | d \in S_\delta^m, \frac{1}{M} \le |d| \le M \right\}}{\left\{ b | b \in S_\beta^n, \frac{1}{M} \le |b| \le M \right\}} \right| = \frac{(\delta - 1)\delta^{m-1}}{(\beta - 1)\beta^{n-1}} \log_\delta \beta. \quad (1)$$

*Proof:* Let $[x]$ denote the greatest integer in $x$. Then the closed interval $[1/M, M]$ may be divided into $2[\log_\beta M]$ disjoint half-open half-closed intervals of the form $[\beta^j, \beta^{j+1})$ and two subintervals of such intervals. Each interval $[\beta^j, \beta^{j+1})$ contains $(\beta - 1)\beta^{n-1}$ distinct members of $S_\beta^n$. Noting that $-b \in S_\beta^n \Leftrightarrow b \in S_\beta^n$ we have for $M \ge 1$,

$$\left| \left\{ b | b \in S_\beta^n, \frac{1}{M} \le |b| \le M \right\} \right|$$

$$= 2 \left| \left\{ b | b \in S_\beta^n, \frac{1}{M} \le b \le M \right\} \right|$$

$$= 2(2\lfloor \log_\beta M \rfloor + \varepsilon)(\beta - 1)\beta^{n-1} \quad \text{where } 0 \le \varepsilon \le 2$$

$$= 2(2 \log_\beta M + \varepsilon')(\beta - 1)\beta^{n-1} \quad \text{where } |\varepsilon'| \le 2,$$

the latter resulting from removal of the greatest integer brackets. Finally,

$$\lim_{M \to \infty} \left| \frac{\left\{ d | d \in S_\delta^m, \frac{1}{M} \le |d| \le M \right\}}{\left\{ b | b \in S_\beta^n, \frac{1}{M} \le |b| \le M \right\}} \right|$$

$$= \lim_{M \to \infty} \frac{2(2 \log_\delta M + \varepsilon_2)(\delta - 1)\delta^{m-1}}{2(2 \log_\beta M + \varepsilon_1)(\beta - 1)\beta^{n-1}} \quad \text{where } |\varepsilon_1|, |\varepsilon_2| \le 2$$

$$= \frac{(\delta - 1)\delta^{m-1}}{(\beta - 1)\beta^{n-1}} \log_\delta \beta.$$

In the folklore on conversion there is an often quoted notion that a decimal digit is equivalent to $\log_2 10 = 3.32 \cdots$ bits. Thus a three-digit decimal system should be only slightly less numerous than a ten-bit binary system. However, if we let $|S_\delta^m/S_\beta^n|$ symbolically denote the limiting ratio given in (1), then $S_{10}^3/S_2^{10} = 0.529 \cdots$, so that there are actually only 53 percent as many real numbers representable with three significant decimal digits as there are real numbers representable with ten significant bits. Furthermore,

the ratio of the number of members of $S_2^3$ to $S_{10}^1$ over the range shown in Fig. 1 was not atypical since $|S_2^3/S_{10}^1| = 1.476 \cdots$, attesting to the fact that there are about 50 percent *more* 3-significant-bit binary numbers than 1-significant-digit decimal numbers, and providing a clear contradiction to the digit = 3.32-bits rule. This anomaly prevails even with more digits, since for large integral $m$ and $n$ chosen such that $10^m/2^n$ approaches unity,

$$\lim_{\substack{m \to \infty \\ 10^m/2^n \to 1}} |S_{10}^m/S_2^n| = \frac{9}{5} \log_{10} 2 = 0.5418 \cdots$$

The results just obtained are not inexplicable on intuitive grounds. The relation "digit = $\log_2 10$ bits" comes from an entropy argument where all different states (values) of the system (digit sequence) are distinguishable. Despite the applicability of this notion to integer and fixed-point number systems, floating-point systems have redundant representations for some numbers, generally resolved by normalization, so that the $\beta^n$ patterns of digits possible for $k$ for each $j$ in $k\beta^j$ do not yield $\beta^n$ new different numbers for each $j$. Furthermore, the collection of normalized numbers corresponding to a fixed power of the base are spread over an interval whose length depends on the base, and both of these conditions are reflected in the final form of (1).

Now one approach for determining a floating-point "equivalent digit formula" is to equate the right-hand side of (1) to unity and solve for $m$ in terms of $n$, $\beta$, and $\delta$. Thus if $S_\delta^m$ is said to be *more dense* than $S_\beta^n$ when $|S_\delta^m/S_\beta^n| > 1$, then from (1), the following occurs.

*Corollary 1:* $S_\delta^m$ is more dense than $S_\beta^n$ if and only if

$$m > n \log_\delta \beta + \log_\delta \left( \frac{\delta(\beta - 1)}{\beta(\delta - 1)} \right) - \log_\delta \log_\delta \beta. \quad (2)$$

Attributing meaning to a nonintegral number of digits, one may propose that

$$m = n \log_\delta \beta + \log_\delta [\delta(\beta - 1)/\beta(\delta - 1)] - \log_\delta \log_\delta \beta \quad (3)$$

is the "equivalent digit formula for floating-point number systems." For binary-decimal conversion we then would have

$$\# \text{ bits} = 3.32 \cdots \times (\# \text{ decimal digits}) - 0.884 \cdots. \quad (4)$$

The variability of spacing of floating-point numbers of a given $S_\beta^n$ is such that the simplified formula (3) does not really provide an adequate comparison of differently based floating-point systems and more attention to local magnitude dependent variability must be considered. In studying the internal structure of $S_\beta^n$, note that every one of its nonzero members will have both a next largest and next smallest neighbor in $S_\beta^n$.

*Definition:* The *successor* $b'$ of $b \in S_\beta^n$, $b \ne 0$, is given by

$$b' = \min \{ d | d > b, d \in S_\beta^n \}, \quad (5)$$

and since distinct members of $S_\beta^n$ have distinct successors, $b$ may then be referred to as the (unique) *predecessor* of $b'$ in $S_\beta^n$.
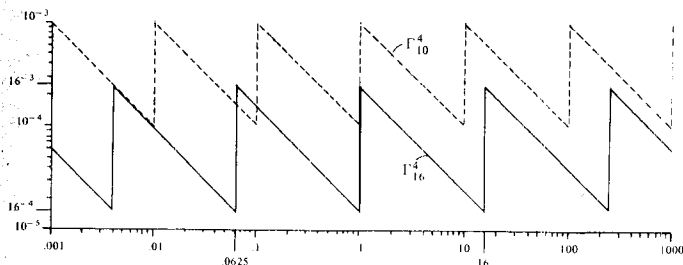
Fig. 2. The gap functions $\Gamma_{10}^4(x)$ and $\Gamma_{16}^4(x)$ for $0.001 \leq x \leq 1000$.

Now the absolute difference, $b' - b$, will grow with $b$ in $S_\beta^n$, however, the relative difference is bounded [1].

*Definition:* The *gap* $\Gamma_\beta^n(x)$ in $S_\beta^n$ at $x \neq 0$ is given by

$$\Gamma_\beta^n(x) =$$
$$\frac{\min\{b|b > x, b \in S_\beta^n\} - \max\{b|b \leq x, b \in S_\beta^n\}}{x} \quad \text{for } x > 0 \quad (6)$$
$$\Gamma_\beta^n(-x) \quad \text{for } x < 0.$$

Specifically then, $\Gamma_\beta^n(a) = (a' - a)/a$ for $0 < a \in S_\beta^n$. From the structure of floating-point number systems it is evident that $\Gamma_\beta^n(\beta x) = \Gamma_\beta^n(x)$, so $\Gamma_\beta^n$ will experience a log periodic behavior. For $1 \leq x < \beta$, the numerator of (6) will have the constant value $\beta^{1-n}$, so that on a log–log scale the gap function appears as a saw tooth function. In Fig. 2, sections of the gap functions $\Gamma_{10}^4$ and $\Gamma_{16}^4$ are illustrated. Note that the variation in the magnitude of the gap function is greater for larger bases.

From Theorem 1 we can calculate that $|S_{16}^4/S_{10}^4| = 5.66 \cdots$, so that overall there are more than five times as many four-significant-digit hexadecimal numbers as four-significant-digit decimal numbers. Yet from the gap functions in Fig. 2, it is apparent that over the interval $(0.0625, 0.1000)$ there are more four-significant-digit decimal numbers than four-significant-digit hexadecimal numbers since $\Gamma_{10}^4(x) < \Gamma_{16}^4(x)$ over that interval.

From the log periodic behavior of $\Gamma_\beta^n$ the following bounds are immediate.

*Theorem 2:* The function $\Gamma_\beta^n$ attains both a minimum and a maximum value over the nonzero members of $S_\beta^n$ given by

$$\min\{\Gamma_\beta^n(b)|b \in S_\beta^n, b \neq 0\} = 1/(\beta^n - 1)$$
$$\max\{\Gamma_\beta^n(b)|b \in S_\beta^n, b \neq 0\} = 1/\beta^{n-1}, \quad (7)$$

and over the nonzero reals the bounds on $\Gamma_\beta^n(x)$ are given by

$$\inf\{\Gamma_\beta^n(x)|x \neq 0\} = 1/\beta^n$$
$$\max\{\Gamma_\beta^n(x)|x \neq 0\} = 1/\beta^{n-1}. \quad (8)$$

Thus the gap function presents a more complete picture of the structure of a floating-point number system than any "equivalent digit" notion, and simplified formulas such as (3) and (4) must be used with extreme caution in any comparison of differently based floating-point number systems.

## III. Conversion Mappings

A conversion procedure determines a specific value in $S_\beta^n$ for each real number $x$. Thus, a conversion process may

be characterized as either a function or a mapping. Formalization of conversion as a mapping appears preferable since it is often useful to invert the question and refer to the set of points which map into a given element of $S_\beta^n$, and this notion is then readily available by considering the inverse mapping. Certainly any conversion mapping of the reals to $S_\beta^n$ should be the identity on $S_\beta^n$ and in addition it is desirable for this mapping to satisfy certain order-preserving properties.

*Definition:* A mapping $M$ of a set $R'$ of real numbers into the reals is

1) *weakly order-preserving* (isotone [6], monotone) *on $R'$* if $x < y \Rightarrow M(x) \leq M(y)$ for all $x, y \in R'$
2) *strongly order-preserving on $R'$* if $x < y \Rightarrow M(x) < M(y)$ for all $x, y \in R'$.

Furthermore, the mapping is said to be *weakly* (*strongly*) *order-preserving* if it is weakly (strongly) order-preserving on the reals.

No conversion mapping of the reals into $S_\beta^n$ can be strongly order-preserving; however, we can expect that a conversion process should at least be weakly order-preserving in addition to being the identity on $S_\beta^n$. These latter two conditions do assure that the inverse image of any $b \in S_\beta^n$ under a conversion mapping is an interval containing $b$. Formally we shall limit our discussion to the rounding and truncation (sometimes called chopping) conversion procedures usually encountered in computerized numeric processing.

*Definition:* The *truncation conversion mapping* $T_\beta^n$ and the *rounding conversion mapping* $R_\beta^n$ of the real numbers into $S_\beta^n$ are defined for all integers $\beta \geq 2$, $n \geq 1$ as follows.

*Truncation Conversion:*

$$T_\beta^n(x) = \begin{cases} \max\{b|b \leq x, b \in S_\beta^n\} & \text{for } x \geq 0 \\ \min\{b|b \geq x, b \in S_\beta^n\} & \text{for } x < 0. \end{cases} \quad (9)$$

*Rounding Conversion:*

$$R_\beta^n(x) = \begin{cases} \min\left\{b\left|\frac{b + b'}{2} > x, b \in S_\beta^n\right.\right\} & \text{for } x > 0 \\ \min\left\{b\left|\frac{b + b'}{2} \geq x, b \in S_\beta^n\right.\right\} & \text{for } x < 0 \\ 0 & \text{for } x = 0. \end{cases} \quad (10)$$

The effects of these conversion mappings in the neighborhood of a power of the base are shown in Fig. 3. Note that the distinctions in the definitions of the mappings of positive and negative values are required to achieve the desired *sign complementary* relations:

$$T_\beta^n(-x) = -T_\beta^n(x), \qquad R_\beta^n(-x) = -R_\beta^n(x). \quad (11)$$

Note that $R_\beta^n((b + b')/2) = b'$ for all positive $b \in S_\beta^n$, so that we have not imposed the additional symmetric rounding condition for midpoints dependent on the "parity" of $b$ that some researchers prefer [7]. Although this refinement could be added without materially affecting our results, we prefer the definition above.
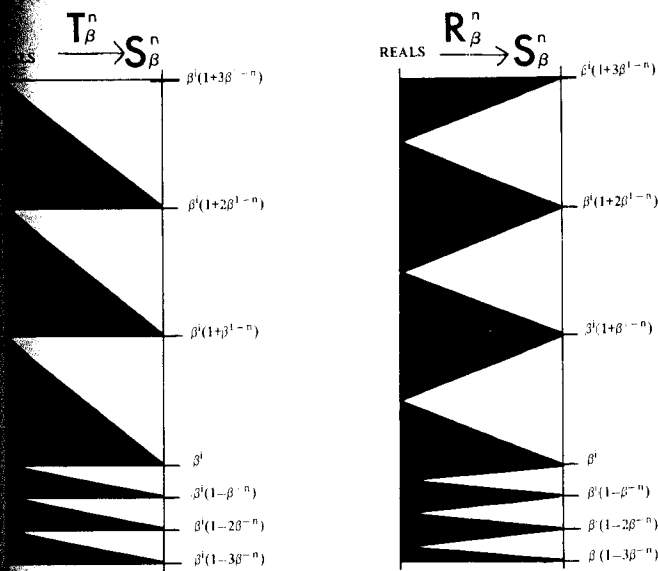
Fig. 4. Sections of the gap functions $\Gamma_{16}^6$ for the 6 (significant)-digit hexadecimal numbers and $\Gamma_8^8$, for the 8 (significant)-digit octal numbers. Hexadecimal and octal are commensurable bases and share a common period of $2^{12}$ on the log scale. Note that the six-digit hexadecimal numbers and the eight-digit octal numbers are not equivalent floating-point systems.



Fig. 3. Conversion of the real numbers to the $n$ (significant) digit base $\beta$ numbers in the neighborhood of a power of the base by truncation conversion $T_\beta^n$ and rounding conversion $R_\beta^n$.

It is evident that $T_\beta^n$ and $R_\beta^n$ both are weakly order-preserving mappings which are identities on $S_\beta^n$ as desired. In practical numeric computation we often need to convert data already expressed in floating-point form to a differently based floating-point form. Thus we are interested in the properties of the restricted mappings $R_\beta^n|S_\delta^m \to S_\beta^n$ and $T_\beta^m|S_\delta^m \to S_\beta^n$.

From consideration of the algorithmic mechanics of conversion, there is evidently a considerable difference between binary–hexadecimal conversion and binary–decimal conversion. Since special relationships between bases do affect the properties of restricted conversion mappings, there is need to characterize this important "commensurability" relationship between bases.

*Definition:* Let $\beta \geq 2$ be a root-free integer, i.e., $\beta$ has no integral $i$th root for any $i$. Then the numbers $\beta, \beta^2, \beta^3, \cdots$ form a *commensurable family* of bases termed the $\beta$-family of bases. Two or more bases belonging to the same commensurable family of bases are *commensurable bases*. Two bases which do not belong to a common commensurable family are termed *incommensurable bases*. Furthermore, two or more significance spaces will be termed *commensurable* when their bases are commensurable.

Thus two, eight, and sixteen are commensurable bases, whereas base ten is incommensurable with any member of the binary family of bases. The root-free condition on $\beta$ in this definition simply assures that each base is in precisely one family.

A useful equivalent characterization of commensurable and incommensurable bases avoiding explicit mention of the respective families to which the bases belong is provided in the following.

*Lemma 1:* The bases $\beta$ and $\delta$ are commensurable if and only if $\beta^i = \delta^j$ for some nonzero integers $i, j$.

*Corollary 2:* The bases $\beta$ and $\delta$ are commensurable if and only if $\log_\delta \beta$ is rational.



Fig. 5. Conversion by rounding of the 3 (significant)-bit binary numbers to the 1 (significant)-digit decimal numbers over the range [1, 1000], indicating that $R_{10}^1|S_2^3$ is neither one-to-one nor onto.

Thus, the gap functions $\Gamma_\beta^n$ and $\Gamma_\delta^m$ plotted on a log–log scale will share a common period when $\beta$ and $\delta$ are commensurable, as in Fig. 4, and otherwise will not (see Fig. 2).

It is readily verified that the conversion by rounding or truncation from a fixed-point number system with $n$ $\beta$-ary digits to the right of the radix point to another fixed-point system with $m$ $\delta$-ary digits to the right of the radix point will be one-to-one if $m \geq n \log_\delta \beta$ and onto if $m \leq n \log_\delta \beta$. Thus a conversion between fixed-point number systems must be either one-to-one or onto; however, a conversion between floating-point number systems need be neither one-to-one nor onto, as seen in Fig. 5. The necessary and sufficient conditions for rounding and truncation conversions between incommensurable significance spaces to be 1) one-to-

one mappings and 2) onto mappings have been determined [2] in the base conversion theorem.

*Theorem 3 (Base Conversion Theorem):* For incommensurable bases $\beta$ and $\delta$, the truncation (rounding) conversion mapping of $S_\beta^n$ to $S_\delta^m$, i.e., $T_\delta^m | S_\beta^n \to S_\delta^m$ $(R_\delta^m | S_\beta^n \to S_\delta^m)$

1) is one-to-one if and only if $\delta^{m-1} \geq \beta^n - 1$

2) is onto if and only if $\beta^{n-1} \geq \delta^m - 1$. $\qquad$ (12)

The details of the proof of this theorem are given in [2]; however, an intuitive understanding for the result can be gleamed from comparing the gap functions (see Fig. 2). Certainly if $\Gamma_\delta^m$ is uniformly less than $\Gamma_\beta^n$, then the mapping of $S_\beta^n$ to $S_\delta^m$ should be one-to-one. Conversely, if the maximum of $\Gamma_\delta^m$ falls above the minimum (restricted to $S_\beta^n$) of $\Gamma_\beta^n$, then from a theorem of Kronecker [8] (that the integer multiples of an irrational number mod 1 are dense in the unit interval) it can be shown that for some $b \in S_\beta^n$, $\Gamma_\delta^m(b) > \Gamma_\beta^n(b)$.

In summary, the essential effect of formulae (12) regarding the conversion of the initial system $S_\beta^n$ to the target system $S_\delta^m$ is that to assure one-to-one conversion a digit must be sacrificed in the target system and to assure onto conversion a digit must be sacrificed in the initial system.

The conditions for one-to-one conversion guarantee another desirable property of the conversion mapping, since it is readily shown that a weakly order-preserving mapping on $R'$ is strongly order-preserving on $R'$ if and only if the mapping is one-to-one from $R'$ to the reals.

*Corollary 3:* For $\beta$ and $\delta$ incommensurable, $T_\delta^m$ and $R_\delta^m$ are each strongly order-preserving on $S_\beta^n$ if and only if $\delta^{m-1} \geq \beta^n - 1$.

It is instructive to consider the implications of the base conversion theorem for decimal-to-binary conversion. It follows that the mapping by rounding or truncation conversion of a decimal-based significance space to a binary-based significance space will be one-to-one (and strongly order-preserving) if and only if

$$\#\text{bits} > 3.32 \cdots \times (\#\text{digits}) + 1 \qquad (13)$$

and onto if and only if

$$\#\text{bits} \leq 3.32 \cdots \times (\#\text{digits}) - 3.32 \cdots. \qquad (14)$$

Relating these inequalities (13), (14) to the membership density formula (1), we conclude that decimal-to-binary floating-point conversion will be 1) one-to-one and strongly order-preserving if and only if the decimal system has less than 0.9 $(\log_{10} 2) = 0.270$ times the membership of the binary system, and 2) onto if and only if the decimal system has more than 18 $(\log_{10} 2) = 5.418$ times the membership of the binary system. Thus conversions between floating-point systems of nearly equal density will be neither one-to-one nor onto.

The properties of decimal–binary conversion are succinctly presented in Fig. 6. The lattice point $n, m$ corresponds to the binary system $S_2^n$ and the decimal system $S_{10}^m$. Lattice points falling to the left of the line $n = (\log_2 10) m + 1 = 3.32m + 1$ correspond to decimal-to-binary conversions which are one-to-one and strongly order-preserving, as well as binary-
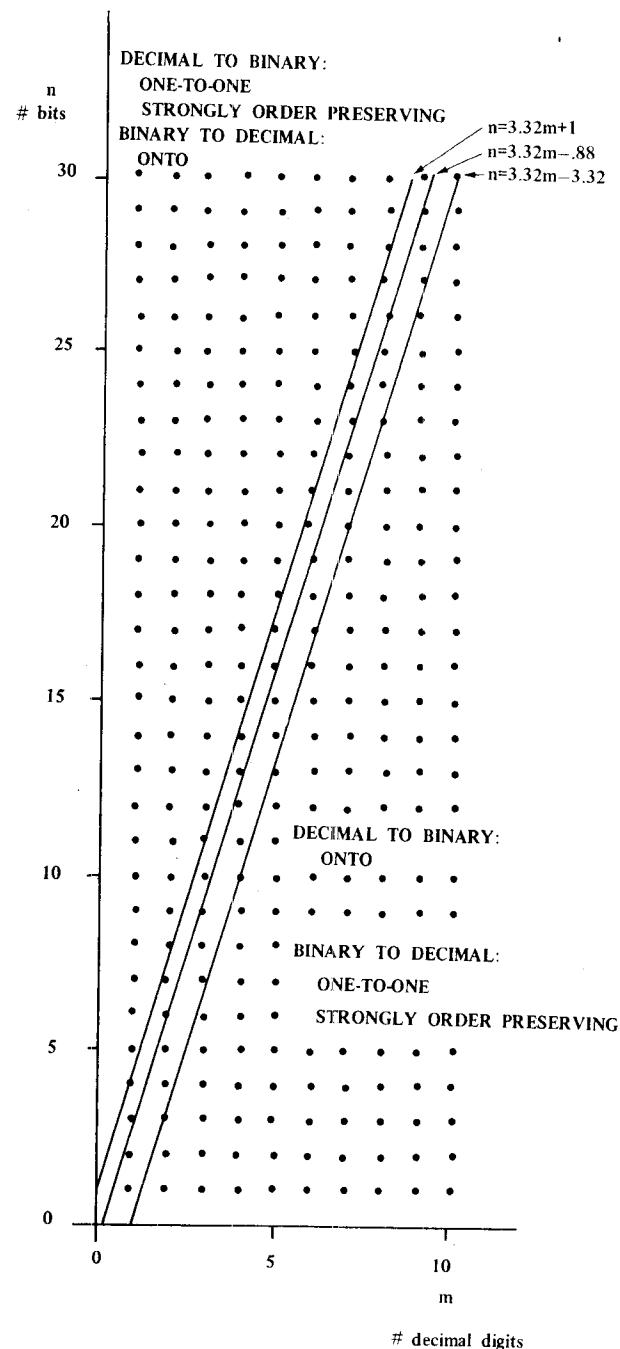


Fig. 6. A summary of the properties of floating-point decimal–binary base conversion.

to-decimal conversions which are onto. Lattice points falling to the right of the line $n = 3.32m - 3.32$ correspond to decimal-to-binary conversions which are onto and to binary-to-decimal conversions which are one-to-one and strongly order-preserving. Lattice points falling between $n = 3.32m + 1$ and $n = 3.32m - 3.32$ correspond to decimal-binary conversions which have none of these three properties. The equal density line $n = 3.32m - 0.88$ separates the lattice points so that those to the left correspond to binary systems which are more dense than the decimal systems, and lattice points to the right correspond to the decimal system being more dense. An increase of one unit on the $n$ axis increases $|S_2^n / S_{10}^m|$ by a factor of 2, and a one unit in-

crease on the $m$ axis decreases this density ratio by a factor of 10, so $|S_2^n/S_{10}^m|$ may be easily estimated by determining the distance of the lattice point $n, m$ from the equal density line.

## IV. COMPOUND CONVERSIONS

The preceding section dealt only with the properties of a single conversion mapping. We have indicated previously that computing environments such as PL/I language programming, multicomputer networks and BCD tape updating on a binary machine present situations where data may be subjected to multiple base conversions during overall job execution. Considerable care must be provided in such mixed-base computing environments to avoid excessive accumulation of error in purportedly "constant" data. To critically analyze this problem, the notion of compound conversion is formally introduced.

*Definition:* For all $n \geq 1$, $\beta \geq 2$,

1) $T_\beta^n$ and $R_\beta^n$ are *1-fold compound conversions through* $S_\beta^n$
2) for $Q$ a $k$-fold compound conversion through $S_{\beta_1}^{n_1}$, $S_{\beta_2}^{n_2}, \cdots, S_{\beta_k}^{n_k}$, $T_\beta^n Q$ and $R_\beta^n Q$ are $(k+1)$-*fold compound conversions through* $S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \cdots, S_{\beta_k}^{n_k}, S_\beta^n$.

Furthermore, $Q$ is a *compound truncation (rounding) conversion* if all the individual conversions composing $Q$ are truncation (rounding) conversions.

Thus $R_\beta^n R_\delta^m$ is a 2-fold compound rounding conversion through $S_\delta^m$, $S_\beta^n$, and $R_\delta^m T_\alpha^i R_\delta^m R_\beta^n$ is a 4-fold compound conversion through $S_\beta^n$, $S_\delta^m$, $S_\alpha^i$, $S_\delta^m$ (see Fig. 7).

A compound conversion is a composition of mappings, and many properties of individual mappings readily carry over to compositions of such mappings. Thus the following important properties of compound conversions are immediate.

*Lemma 2:* If $Q$ is a compound conversion, then $Q(-x) = -Q(x)$ for all $x$.

*Lemma 3:* Compound conversions are weakly order-preserving.

An evident property of truncation conversion is that the magnitude of $T_\beta^n(x)$ is never larger than the magnitude of $x$. Thus truncation conversion performs a contraction of the reals towards zero [see Fig. 3].

*Definition:* The function $M$: Reals→Reals is a *contraction* if $M(x)$ has the same sign as $x$ and $|M(x)| \leq |x|$ for all $x$.

Clearly a composition of contractions is a contraction.

*Lemma 4:* Compound truncation conversions are contractions.

The mappings $R_\beta^n$ and $T_\beta^n$ are identities on their image space $S_\beta^n$; however, $T_2^4 T_{10}^2(1.8) = 1.75$ and $T_2^4 T_{10}^2(1.75) = 1.625$, so that members of the image space of a compound conversion $Q$ are not necessarily carried into themselves by the mapping $Q$. For the compound conversion $Q$ illustrated in Fig. 7, note that $q_1$, $q_5$, and $q_6$ are mapped into themselves by $Q$, whereas $q_2$ and $q_3$ are points of the image space of $Q$ which are not mapped into themselves.

In general, the points mapped into themselves by a compound conversion may be difficult to determine, but they are nonetheless important, since the weakly order-preserv-
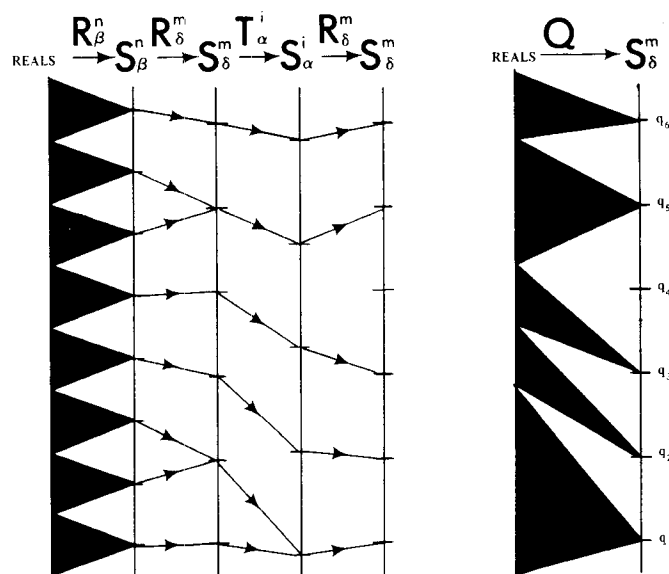


Fig. 7. The composition of a compound conversion from successive conversions. (a) The sequence of conversions $R_\beta^n$, $R_\delta^m$, $T_\alpha^i$, $R_\delta^m$. (b) The 4-fold compound conversion $Q = R_\delta^m T_\alpha^i R_\delta^m R_\beta^n$.

ing property of compound conversions assures us that no point in an interval between two such invariant points of a compound conversion can be mapped outside that interval by that compound conversion.

Formally points $x$ such that $M(x) = x$ are generally referred to [6] as fixed points of the mapping $M$. However, to avoid confusion with the computational notion of fixed-point numbers, we shall refer to fixed points of mappings as invariant points.

*Definition:* Let $M$ be a mapping of the reals to the reals. Then the *invariant set of $M$*, denoted $f(M)$, is given by

$$f(M) = \{x | M(x) = x\}, \tag{15}$$

and for $x \in f(M)$ we say that $x$ is an *invariant point* of the mapping $M$.

Our first result characterizing the invariant sets of compound conversions states that any element common to all significance spaces through which a compound conversion passes is an invariant point of that compound conversion.

*Lemma 5:* If $Q$ is a $k$-fold compound conversion through

$$S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \cdots, S_{\beta_k}^{n_k}, \quad \text{then} \quad \bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \subset f(Q).$$

*Proof:* For any $x \in \cap S_{\beta_i}^{n_i}$, each of the conversions composing $Q$ maps $x$ into itself since $R_{\beta_i}^{n_i}$ and $T_{\beta_i}^{n_i}$ are identities on $S_{\beta_i}^{n_i}$ for $i = 1, \cdots, k$. Thus $Q(x) = x$, and $x \in f(Q)$.

Furthermore, it is now shown that if $Q$ is a compound truncation conversion, then the points in the intersection of the $S_{\beta_i}^{n_i}$ are the only invariant points of the mapping.

*Theorem 4:* If $Q$ is a $k$-fold compound truncation conversion through $S_{\beta_1}^{n_1}, S_{\beta_2}^{n_2}, \cdots, S_{\beta_k}^{n_k}$, then

$$f(Q) = \bigcap_{i=1}^{k} S_{\beta_i}^{n_i}.$$

*Proof:* For $Q = T_{\beta_k}^{n_k} T_{\beta_{k-1}}^{n_{k-1}} \cdots T_{\beta_1}^{n_1}$, by Lemma 5,

$$\bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \subset f(Q).$$

Now assume

$$x \notin \bigcap_{i=1}^{k} S_{\beta_i}^{n_i},$$

and let us show $Q(x) \neq x$. From sign complementarity of $Q$ (Lemma 2), $x > 0$ may be assumed. Let $j$ be some index such that $x \notin S_{\beta_j}^{n_j}$. Compound truncation conversions are contractions and truncation conversion is weakly order-preserving so that

$$x > T_{\beta_j}^{n_j}(x) \geq T_{\beta_j}^{n_j}(T_{\beta_{j-1}}^{n_{j-1}} \cdots T_{\beta_1}^{n_1}(x)) \geq Q(x).$$

Thus $x \neq Q(x)$, proving the theorem.

Lemma 5 and Theorem 4 show the importance of the intersection of significance spaces with regard to determining the invariant points of compound conversion mappings. In turn, the character of the intersection of significance spaces depends in large part on the commensurability or incommensurability of the significance spaces involved. Specifically, the base sixteen is in the binary family, and every hexadecimal number may be easily converted to binary by writing each hexadecimal digit as the appropriate four bits. For $b \in S_{16}^{6}$, the 24-bit binary representation of the six hexadecimal digits of $b$ may have up to three leading zeros, or some 22-bit binary numbers may not be representable in $S_{16}^{6}$, e.g. $(1 + 2^{-21}) \notin S_{16}^{6}$; however, every 21-bit binary number will be contained in $S_{16}^{6}$. In generalizing this notion it is evident that if $\beta = \delta^p$, then each base $\beta$ digit may be represented by $p$ base $\delta$ digits. Thus an $n$-digit base $\beta$ integer yields an $np$-digit base $\delta$ representation of which no more than $p-1$ leading digits may be zero. Setting $m = (n-1)p + 1$, we then have $S_{\delta}^{m} \subset S_{\beta}^{n}$, $S_{\delta}^{m+1} \not\subset S_{\beta}^{n}$, and the following theorem is derived.

*Theorem 5:* Let $\beta_1, \beta_2, \cdots, \beta_k$ be commensurable bases of the $\delta$ family. Then for $m = 1 + \min_{i} \{(n_i - 1) \log_{\delta} \beta_i\}$,

$$S_{\delta}^{m} \subset \bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \quad \text{and} \quad S_{\delta}^{m+1} \not\subset \bigcap_{i=1}^{k} S_{\beta_i}^{n_i}. \tag{16}$$

Actually Theorem 5 can be sharpened if we also consider the different intervals $[\beta^j, \beta^{j+1}]$. Reasoning as above, it can be shown that by letting $p_i = \log_{\delta} \beta_i$, and setting $m = 1 + \min_{i} \{p_i(n_i - 1) + (j \bmod p_i)\}$, then we have

$$S_{\delta}^{m} \cap [\beta^j, \beta^{j+1}] = \bigcap_{i=1}^{k} S_{\beta_i}^{n_i} \cap [\beta^j, \beta^{j+1}]. \tag{17}$$

Furthermore, if $q$ is the minimizing value of $i$ in the above definition of $m$, then also $S_{\delta}^{m} = S_{\beta_q}^{n_q}$ over the interval $[\beta^j, \beta^{j+1}]$.

These observations along with Theorem 4 yield important properties of certain compound conversions. If $Q$ is a compound truncation conversion through commensurable significance spaces, then $Q(x) \in f(Q)$ for all real $x$. Hence $QQ = Q$, and if $Q^*$ is any compound truncation conversion through the same commensurable significance spaces as $Q$

but in any order, then $Q^* = Q$. Furthermore, when $Q = T_{\beta_1}^{n_1} \cdots T_{\beta_k}^{n_k}$ with $\beta_1, \cdots, \beta_k$ in the same commensurable family, then $|x - Q(x)|/x < \max_i \Gamma_{\beta_i}^{n_i}(x)$, so accumulated conversion error is effectively controlled.

Now it will be shown that the intersection of incommensurable significance spaces does not contain any common significance space, for such an intersection has only a *finite number* of elements.

*Theorem 6:* If $\beta$ and $\delta$ are incommensurable, then $S_{\beta}^{n} \cap S_{\delta}^{m}$ has no more than $2(\beta^n - 1)(\delta^m - 1) + 1$ members.

*Proof:* Let $\mathscr{P}$ be the following set of ordered pairs of integers, $\mathscr{P} = \{(k, k^*) \mid |k| < \beta^n, |k^*| < \delta^m \text{ and } kk^* \geq 1 \text{ or } k = k^* = 0\}$. Define a mapping $P: S_{\beta}^{n} \cap S_{\delta}^{m} \to \mathscr{P}$ as follows. $P(0) = (0, 0)$. Suppose $b \in S_{\beta}^{n} \cap S_{\delta}^{m}$, $b \neq 0$. Then $b = k\beta^j$ where $k$ and $j$ may be chosen uniquely such that $|k| < \beta^n$ and $\beta$ does not divide $k$. Similarly, there are unique $k^*, j^*$ such that $b = k^*\delta^{j^*}$ where $|k^*| < \delta^m$ and $\delta$ does not divide $k^*$. In this case $P(b) = (k, k^*)$. Thus $P$ yields the normalized (right-shifted) integer portions of the representations of any element common to $S_{\beta}^{n}$ and $S_{\delta}^{m}$. It is now shown that $P$ is a one-to-one mapping of $S_{\beta}^{n} \cap S_{\delta}^{m}$ into $\mathscr{P}$.

Clearly only zero is mapped into $(0, 0)$, so let $a, b \in S_{\beta}^{n} \cap S_{\delta}^{m}$ be any two nonzero elements where $P(a) = P(b)$. Then there must exist $k, k^* \neq 0$, and $j, j^*, i, i^*$ such that $a = k\beta^j = k^*\delta^{j^*}$, and $b = k\beta^i = k^*\delta^{i^*}$ where $\beta$ does not divide $k$ and $\delta$ does not divide $k^*$. Then $k/k^* = \delta^{j^*}/\beta^j = \delta^{i^*}/\beta^i$, so that $\delta^{j^*-i^*} = \beta^{j-i}$. Then $j^* = i^*$ and $j = i$ since $\beta$ and $\delta$ are incommensurable, so $a = b$ and $P$ is one-to-one. Now $S_{\beta}^{n} \cap S_{\delta}^{m}$ can have no more members than $\mathscr{P}$ which has $2(\beta^n - 1)(\delta^m - 1) + 1$ elements, thus proving the theorem.

Utilizing the fundamental theorem of arithmetic (unique prime decomposition), much more can be said about the members of $S_{\beta}^{n} \cap S_{\delta}^{m}$ for particular $\beta$ and $\delta$. If $\beta$ and $\delta$ are relatively prime, then all members of $S_{\beta}^{n} \cap S_{\delta}^{m}$ are integers, with the largest such integer strictly less than $\beta^n \delta^m$. If the greatest common divisor of the incommensurable $\beta$ and $\delta$ is a prime number (as in the binary–decimal case), then the smallest positive element of $S_{\beta}^{n} \cap S_{\delta}^{m}$ can be shown to be a negative power of that prime. This smallest positive number exactly representable in both systems can be considerably larger than the underflow bound on a typical computer, for example $2^{-10} = 0.0009765625$ is the smallest positive member of $S_{2}^{24} \cap S_{10}^{7}$.

For a compound truncation conversion $Q$ through $S_{\beta_j}^{n_j}$, $j = 1, \cdots, k$, we have shown that $Q(x) = x$ if and only if $x$ is in the intersection of all significance spaces. Furthermore, if at least one pair of the $\beta_j$ are not commensurable, then the intersection is finite and $Q(x) \neq x$ for any positive $x$ lower than the minimum positive element of the intersection. For such an $x$ the compound truncation conversion $Q$ would have $Q(x) > QQ(x) > QQQ(x) > \cdots > Q^{(i)}(x) > \cdots$, and in fact $\lim_{i \to \infty} Q^{(i)}(x) = 0$ since zero is the only finite accumulation point of $S_{\beta_j}^{n_j}$. Thus for all $i$, $Q^i \neq Q^{i-1}$ for any compound truncation conversion $Q$ involving at least two incommensurable bases, in sharp contrast to a compound truncation conversion through commensurable significance spaces where the iterated conversions immediately converged. Practically speaking, the successive updatings of a

BCD tape [4] on a binary machine could cause some of the "constant" floating-point data to be iteratively converted back-and-forth with each updating. If truncation conversion were adhered to as a standard, some of this data could drift lower in value (see Fig. 8), losing all accuracy, with no error indication provided by the system. Thus truncation conversion should be avoided in mixed-base computation unless all bases are in the same commensurable family.

With rounding conversion, the error accumulation upon successive conversions of a datum among incommensurable as well as commensurable bases is much better controlled. For example, it has been shown [9], [3] that with suitably high significance in the intermediate space $S_\delta^m$, the 2-fold compound conversions $R_\beta^n R_\delta^m$ and $R_\beta^n T_\delta^m$ can both reduce to the identity on $S_\beta^n$.

*Theorem 7 (In-and-Out Conversion Theorem):* For $\beta$ and $\delta$ incommensurable,

1) $R_\beta^n R_\delta^m$ is the identity on $S_\beta^n \Leftrightarrow \delta^{m-1} > \beta^n$
2) $R_\beta^n T_\delta^m$ is the identity on $S_\beta^n \Leftrightarrow \delta^{m-1} \geq 2\beta^n - 1$.

Thus from Theorem 7 we see that for certain compound conversions the invariant set will be the whole image space.

*Corollary 4:* For $\beta$ and $\delta$ incommensurable,

$$1) \quad f(R_\beta^n R_\delta^m) = S_\beta^n \Leftrightarrow \delta^{m-1} > \beta^n$$
$$2) \quad f(R_\beta^n T_\delta^m) = S_\beta^n \Leftrightarrow \delta^{m-1} \geq 2\beta^n - 1. \qquad (18)$$

The in-and-out conversion theorem [3] has been stated for the case of incommensurable bases; however, incorporating the techniques of Theorem 5, a more general result encompassing both commensurable and incommensurable bases can be conveniently derived for the 2-fold compound rounding conversion case.

*Corollary 5:* For $\beta, \delta \geq 2$, let $\gamma$ be the greatest common root of $\beta$ and $\delta$ when $\beta$ and $\delta$ are commensurable, and let $\gamma = 1$ otherwise. Then

$$R_\beta^n R_\delta^m \text{ is the identity on } S_\beta^n \Leftrightarrow \gamma \delta^{m-1} \geq \beta^n. \qquad (19)$$

*Proof:* When $\beta$ and $\delta$ are incommensurable, $\gamma = 1$ and $\delta^{m-1} \neq \beta^n$, so (19) follows from Theorem 7. Otherwise when $\beta$ and $\delta$ are commensurable with greatest common root $\gamma$, then $\delta = \gamma^i$, $\beta = \gamma^j$ where $i$ and $j$ are relatively prime. Hence $S_\beta^n \subset S_\gamma^{jn}$ and $S_\gamma^{(m-1)i+1} \subset S_\delta^m$. Now $\gamma \delta^{m-1} \geq \beta^n$ implies that $(m-1)i + 1 \geq jn$, so $S_\beta^n \subset S_\delta^m$ and $R_\beta^n R_\delta^m$ is clearly the identity on $S_\beta^n$. Alternatively, assuming $\gamma \delta^{m-1} < \beta^n$ means that $(m-1)i + 1 < jn$. Now since $i$ and $j$ are relatively prime, there exists a $k$ such that $k = j - 1 \mod j$, $k = 0 \mod i$. But then $S_\beta^n = S_\gamma^{jn}$ over the interval $[\gamma^k, \gamma^{k+1}]$ and also $S_\delta^m = S_\gamma^{(m-1)i+1}$ over the same interval, so that $(m-1)i + 1 < jn$ means $S_\delta^m$ is contained in but not equal to $S_\beta^n$ over the interval $[\gamma^k, \gamma^{k+1}]$. Thus $R_\beta^n R_\delta^m$ cannot be the identity on $S_\beta^n$, completing the corollary.

When the condition $\gamma \delta^{m-1} \geq \beta^n$ is not obtained, then clearly $S_\beta^n \cap S_\delta^m \subset f(R_\beta^n R_\delta^m) \subsetneq S_\beta^n$, and it is of interest to give some alternative properties sufficient to characterize an invariant point of the compound conversion $R_\beta^n R_\delta^m$.

In general, if the image space of the compound conversion
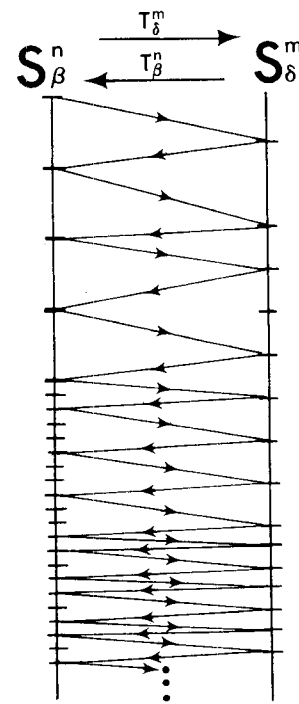


Fig. 8. The possible drift in value of a "constant datum" under iterated truncation conversion between incommensurable significance spaces.

$Q$ were exactly equal to $f(Q)$, then $QQ = Q$, and a desirable situation controlling accumulated error is obtained. If the $k$-fold compound conversion $Q$ ends with a truncation conversion, we have shown that $QQ$ may not equal $Q$. Even if the final conversion of such a $Q$ is a rounding conversion, the image space of $Q$ can contain some points which are not invariant points of $Q$. For example, let $Q = R_3^2 R_2^4$. Since $2^{4-1} = 8 = 3^2 - 1$, by Theorem 3 the mapping $R_3^2 | S_2^4 \to S_3^2$ is onto, so $Q$ covers all of $S_3^2$. However, from Corollary 4, $f(R_3^2 R_2^4) \neq S_3^2$.

Some detailed criteria for determining invariant points of $R_\beta^n R_\delta^m$ will now be considered.

*Lemma 6:* Assume that $b \in S_\beta^n$, $b \neq \pm \beta^i$ for any $i$, and for some $d \in S_\delta^m$, $R_\beta^n(d) = b$. Then $R_\beta^n R_\delta^m(b) = b$, i.e., $b \in f(R_\beta^n R_\delta^m)$.

*Proof:* Let $\beta$ and $\delta$ be commensurable of the $\gamma$ family. Let $b \in S_\beta^n$, $b \neq \pm \beta^i$ for any $i$ and assume $\gamma^j \leq |b| < \gamma^{j+1}$. Over the intervals $[\gamma^j, \gamma^{j+1}]$ and $[-\gamma^j, -\gamma^{j+1}]$ either $S_\beta^n \subset S_\delta^m$ or $S_\delta^m \subset S_\beta^n$, and in either case with $d \in S_\delta^m$, $R_\beta^n(d) = b$ implies that $b \in S_\delta^m \cap S_\beta^n$. Therefore $b \in f(R_\beta^n R_\delta^m)$.

Alternatively, let us assume that $\beta$ and $\delta$ are incommensurable. Let $0 < b' \in S_\beta^n$ have predecessor $b$ and successor $b''$, and assume $b' \neq \beta^i$ for any $i$. Then $b'' - b' = b' - b$. We assume there exists a $d \in S_\delta^m$ such that $R_\beta^n(d) = b'$, so

$$|d - b'| \leq (b' - b)/2. \qquad (20)$$

From the definition of rounding (10), it is evident that $|x - R_\delta^m(x)| = \min_{a \in S_\delta^m} \{|x - a|\}$. Therefore with $x = b'$

$$|b' - R_\delta^m(b')| \leq |b' - d|, \qquad (21)$$

and again with $x = R_\delta^m(b')$

$$|R_\delta^m(b') - R_\beta^n R_\delta^m(b')| \leq |R_\delta^m(b') - b'|, \qquad (22)$$

and finally combining (20)–(22)

$$|b' - R_\beta^n R_\delta^m(b')| \leq b' - b. \tag{23}$$

If (23) were an equality, then $d \neq R_\delta^m(b')$ and (20)–(22) would all be equalities. Therefore $|d - R_\delta^m(b')| = b' - b$, and furthermore there could be no members of $S_\delta^m$ between $d$ and $R_\delta^m(b')$. Also then

$$b' = (d + R_\delta^m(b'))/2 \tag{24}$$

so that $R_\delta^m(b')$ would be the successor of $d$ in $S_\delta^m$. Now any element and its successor in $S_\delta^m$ must differ by an amount $\delta^j$ for some $j$. Similarly, $b' - b = \beta^i$ for some $i$, so that with equality in (23), $\delta^j = \beta^i$. But then $i = j = 0$ since $\delta$ and $\beta$ have been assumed incommensurable. Now $b' - b = \beta^0 = 1$ only if $b$ and $b'$ are consecutive integers, and similarly $d$ and $d' = R_\delta^m(b')$ would be consecutive integers, contradicting (24). Therefore (23) must be a strict inequality, so that $R_\beta^n R_\delta^m(b')$ must equal $b'$. The case for negative elements of $S_\beta^n$ follows from sign complementarity, and, of course, $R_\beta^n R_\delta^m(0) = 0$, completing the lemma.

As a consequence of this lemma, it is now shown that a comparison of the gap functions $\Gamma_\beta^n$ and $\Gamma_\delta^m$ will suffice to determine many of the invariant points of $R_\beta^n R_\delta^m$.

*Corollary 6:* Assume $b \in S_\beta^n$, $b \neq \pm\beta^i$ for any $i$, and $\Gamma_\beta^n(b) \geq \Gamma_\delta^m(b)$. Then $b \in f(R_\beta^n R_\delta^m)$.

*Proof:* If $0 < b' \in S_\beta^n$ with $b' \neq \beta^i$ for any $i$, then $b' - b = b'' - b'$, and the interval mapping into $b'$ under $R_\beta^n$ is the half-open half-closed interval $[(b' - b)/2, (b'' - b')/2]$. Suppose no member of $S_\delta^m$ falls in this interval. Then

$$\max \{d | d < b', d \in S_\delta^m\} \leq (b' - b)/2$$
$$\min \{d | d \geq b', d \in S_\delta^m\} > (b'' - b')/2$$

so that evaluating $\Gamma_\delta^m$ at the real number $b'$ (see (6))

$$\Gamma_\delta^m(b') > (b'' - b')/b' = \Gamma_\beta^n(b').$$

Thus if $\Gamma_\beta^n(b') \geq \Gamma_\delta^m(b')$, then some member of $S_\delta^m$ must map into $b'$ under $R_\beta^n$, and by Lemma 6 $b' \in f(R_\beta^n R_\delta^m)$. The result for negative $b \in S_\beta^n$, $b \neq -\beta^i$, then follows from sign complementarity.

Thus it is readily possible to find sequences of invariant points of $R_\beta^n R_\delta^m$ from a gap function comparison.

*Corollary 7:* For $\beta^i < \delta^j$, assume that $\Gamma_\beta^n \geq \Gamma_\delta^m$ over the open interval $(\beta^i, \min \{\beta^{i+1}, \delta^j\})$. Then

$$(\beta^i, \min \{\beta^{i+1}, \delta^j\}) \cap S_\beta^n \subset f(R_\beta^n R_\delta^m).$$

Referring back to Fig. 2, it is evident from Corollary 7 that all members of $S_{16}^4$ greater than 0.0625 and less than 0.1 are invariant points of $R_{16}^4 R_{10}^4$. Generally not all members of $S_\beta^n$ from an interval $I$ where $\Gamma_\beta^n < \Gamma_\delta^m$ will be members of $f(R_\beta^n R_\delta^m)$, but certainly any point of $R_\beta^n(S_\delta^m)$ other than a power of $\beta$ must be an invariant point of $R_\beta^n R_\delta^m$. Since $R_\beta^n$ restricted to $S_\delta^m \cap I$ is one-to-one to $S_\beta^n$, we may surmise that the number of members of $f(R_\beta^n R_\delta^m)$ in a neighborhood of $x$ is comparable to the lesser of the numbers of members of $S_\beta^n$ and $S_\delta^m$ in that neighborhood. Hence although the members of $f(R_\beta^n R_\delta^m)$ are more erratically spaced, the relative difference between neighboring points of $f(R_\beta^n R_\delta^m)$ will

still be bounded with a bound perhaps larger but not by an order of magnitude than the largest value attained by $\Gamma_\beta^n$ and $\Gamma_\delta^m$.

The preceding lemma and its corollaries do not quite provide the full story about $f(R_\beta^n R_\delta^m)$, since the integral powers of $\beta$ and $\delta$ represent break points in their respective gap functions, and by the preceding theory they must be treated separately. We have already pointed out that the image space of $Q$ need not be identical to $f(Q)$ even for $Q = R_\beta^n R_\delta^m$, so the question of whether the iterates of $Q$, namely $Q$, $QQ$, $QQQ$, $\cdots$, $Q^{(k)}$, $\cdots$, will converge is still unanswered for $Q = R_\beta^n R_\delta^m$. This question is a very practical one since we would like to know if iterated rounding conversion between binary and decimal based systems will allow indefinite drift in the value of a "constant" as did truncation conversion, or if a stable pair of values must be achieved after a fixed number of rounding conversions back and forth. The following theorem is a surprisingly general and reassuring answer to this question.

*Theorem 8 (Iterated Conversion Theorem):* Let $Q = R_\delta^m R_\beta^n$ where $m$, $n \geq 2$. Then $QQQ = QQ$. Furthermore, this result is best possible in the sense that there exists $\beta$, $\delta$, $n$, $m \geq 2$ such that $R_\beta^n QQ \neq R_\beta^n Q$.

*Proof:* For any real $x$, $R_\delta^m R_\beta^n R_\delta^m R_\beta^n(x) = R_\delta^m R_\beta^n(x)$ unless $R_\delta^m R_\beta^n(x) = \pm\delta^i$ for some $i$ by Lemma 6. Similarly, with $Q = R_\delta^m R_\beta^n$, $QQQ(x) = QQ(x)$ unless $QQ(x) = \pm\delta^j$ with $j \neq i$. Assuming $QQQ(x) \neq QQ(x)$, the definition of rounding conversion assures that

$$|R_\delta^m R_\beta^n(x) - R_\beta^n(x)| \geq |R_\beta^n R_\delta^m R_\beta^n(x) - R_\delta^m R_\beta^n(x)|$$
$$\geq |R_\delta^m R_\beta^n R_\delta^m R_\beta^n(x) - R_\beta^n R_\delta^m R_\beta^n(x)|,$$

so that with $|Q(x)| = \delta^i$, $|QQ(x)| = \delta^j \neq \delta^i$,

$$|QQ(x) - Q(x)| \leq 2|R_\delta^m R_\beta^n(x) - R_\beta^n(x)| < \delta^i(1 + \delta^{1-m}) - \delta^i$$

and also

$$|QQ(x) - Q(x)| = |\delta^i - \delta^j| = \delta^i|1 - \delta^{j-i}|,$$

and since $m \geq 2$ by assumption in the theorem.

$$|1 - \delta^{j-i}| < \delta^{1-m} \leq 1/\delta \leq 1/2.$$

Now $|1 - \delta^{j-i}|$ has a positive integer value for $j > i$, and for $j < i$, the smallest value of $|1 - \delta^{j-i}|$ is $1 - 1/\delta \geq 1/2$, and this is a contradiction. Hence $QQQ(x) = QQ(x)$ for all $x$.

The remainder of the theorem demonstrating that $R_\beta^n QQ \neq R_\beta^n Q$ for some $n$, $m$, $\beta$, $\delta \geq 2$ is shown in the example of Fig. 9.

Computer hardware involving the bases 2, 8, and 16 of the binary family and the base 10 is in popular usage. Our results so far discuss accumulated error under compound conversion 1) within a commensurable family and 2) between two incommensurable significance spaces. The modification of our results to cover compound conversion between more than two significance spaces, all from just two different commensurable families, is straightforward, so that our theory does cover the current situations where one might expect to encounter mixed base computation.

Fig. 9. Iterated conversions of $x = 1\ 120\ 000$ by the compound conversion $R_2^4 R_{10}^2$, showing $R_{10}^2 R_2^4 R_{10}^2 R_2^4 R_{10}^2 \neq R_2^2 R_2^4 R_{10}^2$.

If a suitable three-state device were perfected for use in computer hardware, a ternary based floating-point system would undoubtedly be implemented and the possibility of mixed base computation among the bases 2, 3, and 10 would ensue. Our final thoughts will then be concerned with compound conversion through a collection of incommensurable significance spaces.

For the significance spaces $S_{11}^5$, $S_2^{14}$, and $S_5^7$, we have $1\ 960\ 563 = (11\ 1A000)_{11} \in S_{11}^5$, $1\ 960\ 576 = (1\ 11011\ 11010\ 10100\ 00000)_2 \in S_2^{14}$, and $1\ 960\ 500 = (10002\ 14000)_5 \in S_5^7$, and the absolute difference between an element and its successor in this neighborhood is 121, 128, and 125, respectively (see Fig. 10). Thus the iterates of the compound rounding conversion $Q = R_5^7 R_2^{14} R_{11}^5$ in this neighborhood will be different for a number of cycles. As seen in Fig. 10, $Q^{(7)}(1\ 960\ 563) = 1\ 961\ 375$, and $Q^{(k)}(1\ 960\ 563) = 1\ 961\ 500$ for $k \geq 8$, hence $Q^{(k)} \neq Q^{(k-1)}$ for at least all $k \leq 8$. Examples such as this one show that utilizing rounding conversion is not enough to successfully restrict the drift in value of a constant datum under compound conversion. From the generalized result on in-and-out conversion given in Corollary 5, a resolution to the problem of controlling overall error growth in the presence of more than two incommensurable bases by intermediate reconversions to a standard significance space is feasible.

Specifically, let $S_{\beta_i}^{n_i}$, $1 \leq i \leq k$ be a collection of significance spaces representing the different floating-point data formats of a mixed-base computational environment. Suppose we introduce an intermediate space $S_\delta^m$ with the significance $m$ small enough such that $R_\delta^m R_{\beta_i}^{n_i}$ is the identity on $S_\delta^m$ for all $i$. Then let all data introduced into the mixed-base computational environment first be converted by rounding to $S_\delta^m$, and let subsequent conversions from $S_{\beta_i}^{n_i}$ to $S_{\beta_j}^{n_j}$ be preceded by a reconversion to $S_\delta^m$, (i.e., $R_{\beta_j}^{n_j} R_\delta^m | S_{\beta_i}^{n_i} \to S_{\beta_j}^{n_j}$). Note that the



Fig. 10. Iterates of the compound rounding conversion $Q = R_{11}^5 R_2^{14} R_5^7$ showing the drift in value of a "constant datum" under successive conversions.

conversion from $S_{\beta_i}^{n_i}$ to $S_\delta^m$ always regenerates the same value, $R_\delta^m(x)$, in $S_\delta^m$ for an initial datum $x$, since $R_\delta^m R_{\beta_i}^{n_i}$ is the identity on $S_\delta^m$. Thus the value of the initial datum $x$ whenever encountered in $S_{\beta_i}^{n_i}$, even after numerous intermediate conversions, is given by $R_{\beta_i}^{n_i} R_\delta^m(x)$, and this standardization of a constant's value with regard to each $S_{\beta_i}^{n_i}$ provides a highly desirable property for mixed-base computation.

Now the range of possible values achievable in $S_{\beta_i}^{n_i}$ is $R_{\beta_i}^{n_i}(S_\delta^m)$, and it is of course desirable to have as large an $m$ as possible so that the conversion through $S_\delta^m$ does not introduce too much error. Yet it should be kept in mind that if $m$ is chosen so large that one of the $R_\delta^m R_{\beta_i}^{n_i}$ is not the identity on $S_\delta^m$, then conversion error may accumulate and generate a greater overall error than if a smaller $m$ (meaning less initial accuracy) were chosen. These observations can be formalized as an additional corollary to Theorem 7 and Corollary 5.

*Corollary 8:* For $S_{\beta_i}^{n_i}$, $1 \leq i \leq k$, let $\gamma_i$ be the greatest common root of $\delta$ and $\beta_i$ when they are commensurable, and let $\gamma_i$ be unity otherwise. Let

$$m \leq \min_i \{(n_i - 1)\log_\delta \beta_i + \log_\delta \gamma_i\} \qquad (25)$$

and let $Q_i = R_{\beta_i}^{n_i} R_\delta^m$ for $1 \leq i \leq k$. If $Q = Q_j Q'$ for any $1 \leq j \leq k$ where $Q'$ is composed from the mappings $Q_i$, $1 \leq i \leq k$, then $Q = Q_j$.

The fact that $m$ must be bounded from above in (25) in order to guarantee control of accumulated error and avoid situations such as that exhibited in Fig. 10 demonstrates that the phrase "carry more digits" does not always mean that greater overall accuracy will follow, and such cliches should not be used as a substitute for a true understanding of the formal structure of floating-point number systems and base conversion.

## REFERENCES

[1] D. W. Matula, "Base conversion mappings," *1967 Spring Joint Computer Conf.*, AFIPS Proc., vol. 30. Washington, D. C.: Thompson. 1967, pp. 311–318.

[2] ——, "The base conversion theorem," *Proc. Amer. Math. Soc.*, vol. 19, no. 3, pp. 716–723, June 1968.

[3] ——, "In-and-out conversions," *Commun. ACM*, vol. 11, no. 1, pp. 47–50, January 1968.

[4] ——, "Towards an abstract mathematical theory of floating-point arithmetic," *1969 Spring Joint Computer Conf.*, AFIPS Proc., vol. 34. Montvale, N. J.: AFIPS Press, 1969, pp. 765–772.

[5] N. Metropolis and R. L. Ashenhurst, "Radix conversion in an unnormalized arithmetic system," *Math Computation*, vol. 19, pp. 435–441, 1965.

[6] L. Collatz, *Functional Analysis and Numerical Mathematics* (translated by H. Oser). New York: Academic Press, 1966.

[7] M. Urabe, "Roundoff error distribution in fixed-point multiplication and a remark about the rounding rule," *SIAM J. Numer. Anal.*, vol. 5, no. 2, pp. 202–210, 1968.

[8] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 3rd ed. Oxford, England: Clarendon Press, 1954, p. 373.

[9] I. B. Goldberg, "27 bits is not enough for 8-digit accuracy," *Commun. ACM*, vol. 10, pp. 105–106, February 1967.

# The Correspondence Between Methods of Digital Division and Multiplier Recoding Procedures

JAMES E. ROBERTSON, MEMBER, IEEE

*Abstract*—This paper relates previous analyses of the binary SRT division to the theory of multiplier recoding. Since each binary quotient digit has three possible values, the quotient resulting from the SRT division is in recoded form; in this paper it is shown that the recoding is a function of the divisor, and the method for determining the characteristic Boolean function of the recoding is presented. The relationship between the division and the recoding is established by scaling the division in such a way that the scaled "divisor" becomes a constant. Higher radix results are also discussed.

*Index Terms*—Binary arithmetic, division, minimal representations, multiplication, multiplier recoding, redundancy.

## INTRODUCTION

STATISTICAL analyses of the so-called SRT[1][2] method of binary division have been conducted by Freiman [1][2] and Shively [8]. At each recursive step of this division procedure, three alternatives are possible; shift left, add and shift left, or subtract and shift left. One may therefore take the point of view that each quotient digit correspondingly has one of the three values 0, −1, or +1. Thus the division procedure results in a quotient in redundant recoded form. It is the first purpose of this paper to establish the correspondence between the quotient recodings and a class of multiplier recodings.

It is first necessary to review some aspects of the theory of multiplier recoding. For the binary case with recoded digital values of +1, 0, or −1, a recoding can in general be characterized by the choice of two Boolean functions. For the important class of arithmetically symmetric recodings (defined below), the two Boolean functions are duals of one another; hence this class of recodings is characterized by the choice of one Boolean function. A third class of recodings is next defined by restricting the Boolean functions in such a way that each function can be determined by the choice of a single binary numerical parameter in the interval 0 to 1. It is this third class of recodings that corresponds to the quotient recodings of the SRT division.

The next step in establishing the correspondence is that of scaling. It is obvious that the value of the quotient remains unchanged if both the divisor and dividend are multiplied

[1] The earliest published description of a binary division involving redundancy in the representation of the quotient is contained in [4].

[2] Numbers in brackets refer to articles listed in the Reference section.