ISSN 0105-8517

ON MINIMUM WEIGHT BINARY REPRESENTATION OF INTEGERS AND CONTINUED FRACTIONS WITH APPLICATION TO COMPUTER ARITHMETIC *

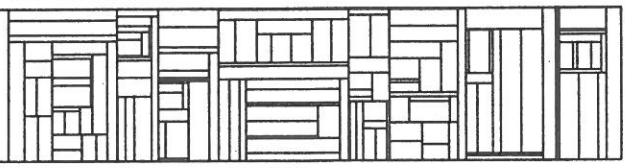
by

David W. Matula
and
Peter Kornerup

DAIMI PB-130
January 1981

* This research was supported in part by the National Science Foundation under grant MCS-8012704.

Computer Science Department
AARHUS UNIVERSITY
Ny Munkegade – DK 8000 Aarhus C – DENMARK
Telephone: 06 – 12 83 55



ON MINIMUM WEIGHT BINARY REPRESENTATION OF INTEGERS AND CONTINUED FRACTIONS WITH APPLICATION TO COMPUTER ARITHMETIC*

by

David W. Matula and Peter Kornerup

Abstract

We develop the concept of minimum weight binary continued fraction representation of a rational number as an extension of minimum weight binary radix representation of an integer. The relation of these representations to the attainment of optimum efficiency in the shift and add or subtract model of binary computer arithmetic is discussed.

* This research was supported in part by the National Science Foundation under Grant MCS-8012704.

I INTRODUCTION AND SUMMARY

Signed digit binary representation allows the digit values of a binary radix polynomial to assume values from the signed digit set $\{-1, 0, 1\}$ rather than the digit set $\{0, 1\}$ of standard binary representation. For example, using $\overline{1}$ to denote the value -1, we have $1\,0\,0\,\overline{1}\,0\,0\,\overline{1}\,0\,1_2 = 1\,1\,0\,1\,1\,1\,0\,1_2 = 221$. Amongst the redundant signed digit representations of an integer n, we are concerned with those representations which have the minimum number, $\omega(n)$, of non zero digits. Thus, from the preceeding example, we note $\omega(221)$ is at most four.

The binary signed continued fraction

$$[a_0, a_1, \ldots, a_m] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_m}}}$$

allows each partial quotient a_i to be a member of the signed binary partial quotient set $\{0, \pm 1, \pm 2, \pm 2^2, \pm 2^3, \ldots\}$, in contrast to standard simple continued fractions where a_0 is any integer and $a_i \geq 1$ for $1 \leq i \leq m-1$, with $a_m \geq 2$. For example, using overbars for negative partial quotients, $[2, 1, \overline{8}, \overline{16}] = [3, 7, 16] = 355/113$.

Amongst the redundant signed continued fraction representations of a rational p/q, we are concerned with those representations which possess the minimum number, $\mu(p/q)$, of non zero partial quotients. From our example, $\mu(355/113)$ is then at most four. It will be shown that $\mu(p/1) =$

$\omega(p)$, so μ: Rationals → Int provides an extension to the rationals of ω: Int → Int, where μ and ω are termed the minimum weight functions over the rationals and over the integers, respectively.

A considerable literature has been implicitly developed for the minimum weight function over the integers incidental to the design of efficient multiplication and division algorithms for binary computers [N56, R58, T58, R60, F61, M61, M62, R70]. Note that in the multiplication of two binary represented integers by an iterative shift and add or subtract procedure, it is the number of non zero digits in the signed digit binary representation of the multiplier that determines the number of add and subtract operations in computing the product.

As a model for standard binary multiplication architecture it is reasonable to assume the add and subtract execution times to be equivalent and more costly than the shifts. Furthermore, the total number of shifts is determined by the position of the leading digit of the multiplier rather than the composition of individual digit values, and so is not a significant variable. For the shift and add or subtract model of multiplication with efficiency measured in terms of the total number of adds and subtracts, the minimum weight binary representation of the multiplier then achieves optimum speed. Detailed investigations of the shift and add or subtract model of the division of binary represented integers by several authors [R58, T58, F61, WL61, M62] has culminated in the result that the quotient, which can be determined by a number of add or subtract operations equal to the number of signed non zero digits in the resulting

representation of the quotient, can always be obtained in one of its minimum weight forms. Thus the minimum weight representation of the resulting quotient also determines the optimum speed in the shift and add or subtract model of division of binary represented integers. Since binary floating point numbers may simply be considered as integers scaled by powers of two, these same optimal multiplication and division results also pertain to binary floating point arithmetic.

Reitwiesner pursues an explicit detailed investigation of minimum weight binary representation in [R60]. He shows that a unique canonical minimum weight binary representation can be characterized by the property that no two successive digits are both non zero, and gives a simple digit by digit right to left conversion procedure from standard to canonical minimum weight binary representation. From analysis of the algorithm he then shows that an assumed average density of one-half for the non zero digits in the standard binary representation yields an average of one-third non zero digits in the minimum weight binary representation. To see that this average is asymptotically achievable consider the following procedure for determining a signed digit binary representation of a real number $\frac{1}{2} < \alpha \leq 1$. Choose a digit $b \in \{\frac{1}{2}, 1\}$ so as to minimize $|b - \alpha|$, with say $b = 1$ for $\alpha = 3/4$. Then multiply $b - \alpha$ by the appropriate power of two, termed the shift length k, to achieve $\frac{1}{2} < 2^k |b-\alpha| \leq 1$, and repeat the cycle. Each cycle computes one non zero digit out of a number of digits(of the binary fraction for α) equal to the average shift length per cycle. If α is chosen uniform on $(\frac{1}{2}, 1]$, then $2^k |b-\alpha|$ is also uniform on $(\frac{1}{2}, 1]$ and the average shift

length is $\frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 3 + \frac{1}{8} \cdot 4 + \ldots = 3$, so the asymptotic average density of non zero digits in such a signed digit representation of α is 1/3.

In Section II we develop properties of the minimum weight function ω: Int → Int in a straightforward number theoretic manner primarily to provide a self contained foundation for the subsequent study of the minimum weight function over the rationals. Our main original contribution of this section is an exact formula for $\Omega(2^k) = \sum_{i=1}^{2^k} \omega(i)$, from which we obtain that the average minimum weight of the integers over the range $0 \leq n \leq 2^k - 1$ is $\frac{1}{3}k + \frac{4}{9} + \frac{(-1)^k - 9}{18 \cdot 2^k}$.

Our interest in minimum weight binary signed continued fractions is likewise motivated by their importance in determining the optimum speed in terms of the number of add and subtract operations in the shift and add or subtract model of binary computer arithmetic. In this case we are specifically interested in the multiplication, division, addition and subtraction of fractions, with the application relating to efficient computer arithmetic unit design for fixed and floating slash arithmetic [MK80].

Essential to efficient computer arithmetic design for such slash arithmetic is the fact that the truncated signed continued fraction $\frac{p_i}{q_i} = [a_0, a_1, a_2, \ldots, a_i]$ of the binary signed continued fraction $[a_0, a_1, a_2, \ldots, a_m]$ can be recursively computed forwardly for $i = 0, 1, 2, \ldots, m$ from the equations

$$
\left.
\begin{aligned}
p_{-2} &= 0 \\
q_{-2} &= 1 \\
p_{-1} &= 1 \\
q_{-1} &= 0 \\
p_i &= a_i\ p_{i-1} + p_{i-2} \\
q_i &= a_i\ q_{i-1} + q_{i-2}
\end{aligned}
\right\} \quad \text{for } 0 \leq i \leq m.
$$

Thus for a_i a power of two, each cycle of the recursion is simply a standard shift and add for the numerators and denominators, which may be computed in parallel in computer hardware. Furthermore, by seeding p_{-1} with the value r and q_{-2} with the value s, we may then determine the product $\frac{r}{s} \times \frac{p_m}{q_m}$ in a number of addition and subtraction operations given by the number of non zero partial quotients in the binary signed continued fraction $[a_0, a_1, \ldots, a_m]$. The following illustrates the computation of $\frac{r}{s} \times [2, 1, \overline{8}, \overline{16}] = \frac{355r}{113s}$.

i	-2	-1	0	1	2	3
a_i			2	1	$\overline{8}$	$\overline{16}$
rp_i	0	r	2r	3r	-22r	355r
sq_i	s	0	s	s	-7s	113s

Our purpose in exploring the feasibility of slash arithmetic is to support reasonably efficient approximate real arithmetic as well as exact rational arithmetic. To this end it may be noted that the preceeding example computes an approximation to $\frac{r}{s} \times \pi$ with relative error less than 1×10^{-7} using only four shift and add operation cycles.

The preceeding fraction multiplication procedure employing continued fraction representation for one of the arguments may be extended to division by simply reciprocating r/s. We may similarly compute [KM81] addition or subtraction of fractions by setting $p_{-2} = \pm r$, $p_{-1} = s$, $q_{-2} = s$ and $q_{-1} = 0$, where then we obtain the resulting fraction $\frac{t}{u} = \frac{p_m}{q_m} \pm \frac{r}{s}$. Related interesting observations on the possibilities for exact computation utilizing rationals with both arguments and results in their continued fraction form have been given by Gosper [G80].

We have found no previous treatment of minimum weight binary continued fractions in the literature. As a foundation for the investigation of minimum weight signed binary continued fractions in Section III, we initially define the minimum weight function μ: Rationals → Int in terms of the minimum weights of the partial quotients of the signed continued fraction $[a_0, a_1, \ldots, a_m]$ where the partial quotients are arbitrary integers. Several useful value preserving transformation rules for signed continued fractions are derived expanding on some observations of Knuth [K69, p. 334-336]. A minimality criteria for signed continued fractions is introduced and the transformation rules are utilized to exclude certain partial quotient subsequences from any such "minimal" continued fraction, whereby we then obtain a recurrance equation for computation of $\mu(\frac{p}{q})$. These results are then related to binary signed continued fractions where we are able, however, to provide only a partial solution to the goal of determining an algorithm that recursively generates the successive partial quotients of a minimum weight binary continued fraction.

Noting the sign and magnitude inversion symmetry of $\mu(\frac{p}{q})$, we need only consider $\frac{p}{q} \geq 1$. Specifically we show for $\frac{p}{q} \geq 4$ with 2^k the closest power of two from $\frac{p}{q}$, that $\mu(\frac{p}{q}) = 1 + \mu(2^k - \frac{p}{q})$. In fact for $\frac{23}{32}\, 2^{k-1} \leq \frac{p}{q} \leq \frac{25}{32}\, 2^{k-1}$, both 2^{k-1} and 2^k can each be the leading term of some minimum weight binary continued fraction for $\frac{p}{q}$. This provides useful redundancy that can be used in minimizing digit lookahead in obtaining the continued fraction for $\frac{p}{q}$ in a manner similar to the extended non restoring division procedure analysed by Wilson, Ledley and Metze [WL61, M62].

For $1 \leq \frac{p}{q} \leq 4$, we exhibit certain regions where the closest power of two can not be the leading term of any minimum weight binary continued fraction. Determination of the appropriate leading term becomes progressively more difficult over successively smaller nested subregions within the interval (1, 4). However, these subregions have progressively smaller measure, so close approximation of the "average" value of $\mu(\frac{p}{q})$ is possible.

Binary continued fraction representations of π and e are specifically investigated. A binary continued fraction for e is derived and shown to yield a succession of approximations to e reducing the relative approximation error by an average factor of about 64 (6 more bits of accuracy) for every non zero binary partial quotient.

In Section IV we tabulate some numerical results for $\mu(\frac{p}{q})$ and the average value given by $\frac{1}{n^2} \sum_{1 \leq p, q \leq n} \mu(\frac{p}{q})$. Extrapolating on these results we conclude with some observations, the foremost of which is that the GCD(p, q) can be computed with an average of only about 5% to 10% more shift and add cycles in this model than the average number of division cycles in the standard Euclidian algorithm.

II MINIMUM WEIGHT BINARY REPRESENTATION OF INTEGERS

A signed digit binary representation of an integer n is a binary radix polynomial $P = b_m 2^m + b_{m-1} 2^{m-1} + \ldots + b_1 2^1 + b_0$ of value n where all b_i are members of the signed digit set $\{-1, 0, 1\}$. The number of non zero digits of P is termed the weight of P. A signed digit binary representation P is a minimum weight representation of the integer n if P has the minimum weight, denoted by $\omega(n)$, over all signed digit binary representations of n. The resulting minimum weight function ω: Int → Int has rather erratic behavior with n, and the following triangle inequality will be useful in analysing the behavior of ω.

Lemma 1 For any integers i, j,

$$\omega(i+j) \leq \omega(i) + \omega(j), \qquad (1)$$

where the inequality is strict whenever i and j are both odd integers.

Proof Let P_i and P_j be minimum weight representations of i and j, respectively. Let C and P be the binary radix polynomials of carries and place values, respectively, determined by one cycle (no carry propagation) of the addition of P_i and P_j using the addition table of Figure 1 in parallel for each digit position.

Digit Values	$\bar{1}$	0	1
$\bar{1}$	$\bar{1}0$	$0\bar{1}$	00
0	$0\bar{1}$	00	01
1	00	01	10

Figure 1: Binary addition table giving (carry, place) values where the overbar denotes a negative digit value.

From the addition table note that the digits in the k'th places of P_i and P_j determine the k'th digit of P and the (k+1)'th digit of C in a manner preserving or decreasing the number of non zero digits for each k. Thus the sum of the weights of C and P is at most $\omega(i)+\omega(j)$, and will be strictly less if both P_i and P_j have a non zero digit in the k'th place for at least one value of k. The addition process may be iterated until the carry polynomial is zero, in which case the corresponding place polynomial then has value i+j and weight at most $\omega(i)+\omega(j)$, verifying (1). Furthermore, if i and j are both odd, then P_i and P_j must both have non zero unit position digits, so then inequality (1) must be strict as previously noted.

□

Corollary 1.1: For any integer k,

$$\omega(2k) \leq \omega(2k \pm 1) \leq \omega(2k) + 1.$$

The minimum weight function may be computed recursively as shown in the following lemma.

Lemma 2 For any integer n,

$$\omega(n) = \begin{cases} 0 & \text{for } n = 0 \\ \omega(k) & \text{for } n = 2k, \\ \omega(k)+1 & \text{for } n = 4k \pm 1. \end{cases} \qquad (2)$$

Proof The value $\omega(0) = 0$ is immediate. Then note that any signed digit binary representation of the even number $n = 2k$ must have the unit place digit equal to zero. Thus $P = b_m 2^m + b_{m-1} 2^{m-1} + \ldots + b_1 2^1 + b_0$ has value $n = 2k$ if and only if $b_0 = 0$ and $P' = P/2 = b_m 2^{m-1} + b_{m-1} 2^{m-2} + \ldots + b_2 2^1 + b_1$ has value k, where since P and P' have equal weight, then $\omega(n) = \omega(k)$.

From Corollary 1.1 with $u = \pm 1$, we obtain $\omega(4k) \leq \omega(4k+u) \leq \omega(4k)+1$, which by the preceeding result gives $\omega(k) \leq \omega(4k+u) \leq \omega(k)+1$. Let $P = b_m 2^m + \ldots + b_1 2^1 + b_0$ be a minimum weight representation of $4k+u$. Then $b_0 \equiv u \bmod 2$, so either $b_0 = u$ or $b_0 = -u$. If $b_0 = u$, then $P' = P - b_0 = b_m 2^m + \ldots + b_1 2^1 + 0$ has value $4k$ and weight one less than P, so $\omega(4k+u) = \omega(k)+1$. Otherwise $b_0 = -u$, so then $P' = P - b_0$ has value $4k + 2u$ and weight one less than P. But then using $\omega(2j) = \omega(j)$ and Corollary 1.1, we have $\omega(4k+u) \geq 1 + \omega(4k+2u) = 1 + \omega(2k+u) \geq 1 + \omega(2k) = 1 + \omega(k)$, and the lemma is proved.

□

In his study of minimum weight binary representation, Reitwiesner [R60] characterized a "canonical" minimum weight representation by the property that no two consecutive digits are both non zero. He showed that a unique canonical minimum weight representation exists for every integer, and then

gave an efficient right to left digit sequential conversion procedure to obtain the canonical minimum weight representation from the standard binary representation. From Lemma 2 we obtain the following algorithm for determining a minimum weight representation of any integer. The representation is readily seen to have the property that no two consecutive digits are both non zero, and hence is the canonical minimum weight representation.

ALGORITHM MINREP(i)

{ For any integer i, this recursive algorithm determines the canonical minimum weight signed digit binary representation of i in digit string form. $\overline{1}$ denotes the digit value -1 and & denotes the string concatenation operation.}

```
begin if i = 0 then MINREP := '0' else
      case i mod 4 of
           0, 2:  MINREP := MINREP(i/2) & '0'
              1:  MINREP := MINREP((i-1)/4) & '0' & '1'
              3:  MINREP := MINREP((i+1)/4) & '0' & '1̄'
      end
end
```

If we assume that the value of i is available in standard binary representation for ALGORITHM MINREP(i), then i/2, (i-1)/4, and (i+1)/4 can be efficiently computed with shifts and carries yielding essentially the conversion algorithm of Reitwiesner [R60]. To see that the canonical representation is

unique, assume two distinct minimum weight representations of some integer j both have no adjacent non zero digits. Without loss of generality assume the representations differ in the units position, hence j is odd, so then the two low order terms of the two distinct representations are $0 * 2^1 + 1$ and $0 * 2^1 + \overline{1}$. But then $1 \equiv j \bmod 4$ and $-1 \equiv j \bmod 4$, a contradiction. Hence the canonical representation is unique.

From Lemma 2 and ALGORITHM MINREP(i) we obtain Table 1.

n	$\omega(n)$	MINREP(n)	n	$\omega(n)$	MINREP(n)
0	0	0	16	1	1 0 0 0 0
1	1	1	17	2	1 0 0 0 1
2	1	1 0	18	2	1 0 0 1 0
3	2	1 0 $\overline{1}$	19	3	1 0 1 0 $\overline{1}$
4	1	1 0 0	20	2	1 0 1 0 0
5	2	1 0 1	21	3	1 0 1 0 1
6	2	1 0 $\overline{1}$ 0	22	3	1 0 $\overline{1}$ 0 $\overline{1}$ 0
7	2	1 0 0 $\overline{1}$	23	3	1 0 $\overline{1}$ 0 0 $\overline{1}$
8	1	1 0 0 0	24	2	1 0 $\overline{1}$ 0 0 0
9	2	1 0 0 1	25	3	1 0 $\overline{1}$ 0 0 1
10	2	1 0 1 0	26	3	1 0 $\overline{1}$ 0 1 0
11	3	1 0 $\overline{1}$ 0 $\overline{1}$	27	3	1 0 0 $\overline{1}$ 0 $\overline{1}$
12	2	1 0 $\overline{1}$ 0 0	28	2	1 0 0 $\overline{1}$ 0 0
13	3	1 0 $\overline{1}$ 0 1	29	3	1 0 0 $\overline{1}$ 0 1
14	2	1 0 0 $\overline{1}$ 0	30	2	1 0 0 0 $\overline{1}$ 0
15	2	1 0 0 0 $\overline{1}$	31	2	1 0 0 0 0 $\overline{1}$

Table 1: The minimum weight $\omega(n)$ and canonical minimum weight representation of n for $n = 0, 1, \ldots, 31$.

In order to obtain a better understanding of the rather erratic function $\omega(n)$, we shall describe its extremal and average behavior for increasing n. Since $\omega(2^i) = 1$ for all $i \geq 0$, lim inf $\omega(n) = 1$. Now lim sup $\omega(n)$ diverges, so let $n_j = \min \{ n \mid \omega(n) = j, n \geq 0 \}$, and note that the canonical minimum weight representation of n_j must have a leading term of size at least $2^{2(j-1)}$. The smallest positive canonical representation with this leading term is

$$2^{2(j-1)} - \sum_{i=0}^{j-2} 2^{2i} = \lceil 2^{2j-1}/3 \rceil .$$

Thus $n_j = \lceil 2^{2j-1}/3 \rceil$ for $j \geq 1$, where for example $n_3 = \lceil 2^5/3 \rceil = 11 = 1\,0\,\bar{1}\,0\,\bar{1}_2$, and $n_5 = \lceil 2^9/3 \rceil = 171 = 1\,0\,\bar{1}\,0\,\bar{1}\,0\,\bar{1}\,0\,\bar{1}_2$. From the formula for n_j we immediately obtain the following extremal result for the behavior of ω.

Lemma 3 lim sup $\omega(n)/\log_2 n = 1/2$.

Certain patterns and symmetries are observed in the values of $\omega(n)$ in Table 1 and the next two lemmas confirm specific patterns for the whole range of ω.

Lemma 4 For any $k \geq 0$,

$$\omega(2^k + i) = \begin{cases} 1 + \omega(i) & \text{for } 0 \leq i < (\frac{2}{3})2^k, \\ \omega(i) & \text{for } (\frac{2}{3})2^k < i \leq 2^k. \end{cases} \tag{3}$$

Proof As a basis for induction note from Table 1 that (3) holds for $k \leq 4$. Let $k \geq 5$, and by induction assume (3) holds for all exponent values through $k-1$. For $0 \leq i \leq 2^k$, using Lemma 2 and the induction assumption,

(i) for $i \equiv 0 \bmod 2$,

$$\omega(2^k+i) = \omega(2^{k-1} + i/2)$$

$$= \begin{cases} 1 + \omega(i/2) & \text{for } 0 \leq i/2 < (\frac{2}{3})2^{k-1} \\ \omega(i/2) & \text{for } (\frac{2}{3})2^{k-1} \leq i/2 \leq 2^{k-1} \end{cases}$$

$$= \begin{cases} 1 + \omega(i) & \text{for } 0 \leq i < (\frac{2}{3})2^k \\ \omega(i) & \text{for } (\frac{2}{3})2^k \leq i \leq 2^k, \end{cases}$$

(ii) for $i = 4j+u$ with $u = \pm 1$, noting $j < (\frac{2}{3})2^{k-2}$ implies $j \leq (\frac{2}{3})2^{k-2} - \frac{1}{3}$,

$$\omega(2^k+i) = \omega(4(2^{k-2}+j)+u) = 1 + \omega(2^{k-2}+j)$$

$$= \begin{cases} 2 + \omega(j) & \text{for } 0 \leq j \leq (\frac{2}{3})2^{k-2} - \frac{1}{3} \\ 1 + \omega(j) & \text{for } (\frac{2}{3})2^{k-2} + \frac{1}{3} \leq j \leq 2^{k-2} \end{cases}$$

$$= \begin{cases} 1 + \omega(4j+u) & \text{for } 0 \leq 4j \leq (\frac{2}{3})2^k - \frac{4}{3} \\ \omega(4j+u) & \text{for } (\frac{2}{3})2^k + \frac{4}{3} \leq 4j \leq 2^k \end{cases}$$

$$= \begin{cases} 1 + \omega(i) & \text{for } 0 \leq i < (\frac{2}{3})2^k \\ \omega(i) & \text{for } (\frac{2}{3})2^k \leq i \leq 2^k \end{cases}$$

and (3) holds for all k by induction.

□

We also state the following noting that the proof follows by induction in the same manner as for Lemma 4.

Lemma 5 For any $k \geq 0$,

$$\omega(2^k+i) = \omega(2^{k+1} - i) \quad \text{for } 0 \leq i \leq 2^k. \tag{4}$$

From Lemmas 4 and 5 note that if $2^k \leq n \leq 2^{k+1}$, then the leading term of a minimum weight representation of n must be either 2^k or 2^{k+1}. Furthermore if $2^k \leq n < (\frac{4}{3})2^k$, then the leading term must be 2^k; if $(\frac{5}{3})2^k < n \leq 2^{k+1}$, then the leading term must be 2^{k+1}; and if $(\frac{4}{3})2^k < n < (\frac{5}{3})2^k$, then the leading term can be either 2^k or 2^{k+1}. The following algorithm utilizes these observations to provide all minimum weight representations of n.

ALGORITHM ALL(n, k)

{This algorithm generates all k digit minimum weight representations of an integer n for $|n| < \frac{4}{3}\,2^{k-1}$ in digit string form. COMPL denotes sign inversion of all digits of all members of a set of strings, & denotes the string concatenation operation, and $\emptyset$ denotes the empty set.}

```
begin if k = 0 then ALL := ∅
      else if n < 0 then ALL := COMPL(ALL(-n, k))
      else if (5/3) 2^(k-2) < n < (4/3) 2^(k-1) then ALL := '1' & ALL(n-2^(k-1), k-1)
      else if (4/3) 2^(k-2) < n < (5/3) 2^(k-2) then
           ALL := '1' & '0' & ALL(n-2^(k-1), k-2) ∪ '0' & '1' & ALL(n-2^(k-2), k-2)
      else if 0 ≤ n < (4/3) 2^(k-2) then ALL := '0' & ALL(n, k-1)
end
```

If n is given in standard binary representation, then inspection of the two most significant bits of n allows the determination that either $2^k \leq n < \frac{3}{2}\,2^k$ or $\frac{3}{2}\,2^k \leq n < 2^{k+1}$. From the preceeding observations note that by associating as leading terms either 2^k or 2^{k+1} with these two conditions, respectively, we may then also readily obtain a left to right digit by digit conversion algorithm from standard binary to a minimum weight binary (but not necessarily canonical) form.

The cumulative sum and cumulative average value of $\omega(n)$ as n increases are now shown to have more stable behavior. Let $\Omega(n) = \sum_{i=1}^{n} \omega(i)$ for any $n \geq 0$.

Theorem 6 For any $k \geq 0$,

$$\Omega(2^k) = \frac{1}{3} k 2^k + \frac{4}{9} 2^k + \frac{1}{2} + \frac{(-1)^k}{18} \quad . \tag{5}$$

Proof As a basis for induction by direct evaluation (5) is seen to hold for $k = 0, 1$. Assume (5) holds for all exponents through $k-1$ for a given $k \geq 2$. From Lemma 4 and the induction assumption,

$$\Omega(2^k) = 2\,\Omega(2^{k-1}) + \lfloor \frac{2^k}{3} \rfloor$$

$$= 2\left(\frac{k-1}{3}\, 2^{k-1} + \frac{4}{9}\, 2^{k-1} + \frac{1}{2} + \frac{(-1)^{k-1}}{18}\right) + \lfloor \frac{2^k}{3} \rfloor$$

$$= \frac{k}{3}\, 2^k + \frac{4}{9}\, 2^k + \lfloor \frac{2^k}{3} \rfloor - \frac{2^k}{3} + 1 + \frac{(-1)^{k-1}}{9},$$

and by considering k even and odd separately we obtain (5) in both cases, and the theorem then follows by induction.

□

From Theorem 6 we are able to make an (exact) comparison of the average weights of standard binary and minimum weight binary representation for the k-bit integers $\{0, 1, 2, \ldots, 2^k - 1\}$ which are of particular importance for number systems and arithmetic employing binary representation.

Corollary 6.1 The average weight of the binary representation of the k-bit integers $\{0, 1, 2, \ldots, 2^k-1\}$ for any k is given by

(i) $\frac{1}{2}k$ for standard binary representation,

(ii) $\frac{1}{3}k + \frac{4}{9} + (\frac{(-1)^k-9}{18})\frac{1}{2^k}$ for minimum weight binary representation.

Proof Note that the sum of the weights of the unique standard binary representations of i and 2^k-1-i is k for any $i \in \{0, 1, 2, \ldots, 2^k-1\}$, so the average of $\frac{1}{2}k$ is obtained for standard binary representation. Since $\omega(0) = 0$, the average minimum weight binary representation over $\{0, 1, 2, \ldots, 2^k-1\}$ is given by $\Omega(2^k-1)/2^k = (\Omega(2^k)-1)/2^k$, so expression (ii) follows from (5).

□

From Lemma 4 and Theorem 6 it is also possible to determine values of $\Omega(n)$ for n not a power of 2, where the results are simplest if the binary expansion of n has relatively few terms. For example, for $k \geq 1$,

$$\begin{aligned}\Omega(\tfrac{3}{2}2^k) &= \Omega(2^k) + \Omega(2^{k-1}) + 2^{k-1} \\ &= 3\Omega(2^{k-1}) + \lfloor \tfrac{5}{3}2^{k-1} \rfloor\end{aligned}$$

and substitution of (5) then would provide an exact formula for $\Omega(\frac{3}{2}2^k)$ for all $k \geq 1$. More simply we note from the substitution that asymptotically in k,

$$\Omega(\tfrac{3}{2}2^k) = \tfrac{1}{3}(\tfrac{3}{2}2^k)\log_2(\tfrac{3}{2}2^k) + \Theta(\tfrac{3}{2}2^k).$$

Utilizing the same procedure for any constant of the form $b = i/2^j$ where $2^{j-1} \leq i < 2^j$ (i.e. a normalized binary fraction), we may similarly obtain asymptotically in k,

$$\Omega(b2^k) = \tfrac{1}{3}(b2^k)\log_2(b2^k) + O(b2^k). \tag{6}$$

From (6) and the fact that $\Omega(n)$ is monotone increasing, we note that $\Omega(n)/(n \log_2 n)$ must converge and so obtain the following corollary.

Corollary 6.2 $$\lim_{n\to\infty} \Omega(n)/(n \log_2 n) = 1/3. \tag{7}$$

III MINIMUM WEIGHT CONTINUED FRACTIONS

The notation $[a_0, a_1, \ldots, a_m]$ shall denote the signed continued fraction

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_m}}}$$

where the partial quotients a_i are arbitrary integers except that $[a_i, a_{i+1}, \ldots, a_m]$ is non zero for $0 \leq i \leq m$ when $m \geq 1$. The latter condition assures that the value of any signed continued fraction is a finite rational number, where furthermore $\frac{p}{q} = [a_0, a_1, a_2, \ldots, a_m] = a_0 + \frac{1}{[a_1, a_2, \ldots, a_m]}$. The standard continued fractions with $a_i \geq 1$ for $1 \leq i \leq m$ are contained within the set of signed continued fractions, so for every rational $\frac{p}{q}$ at least one signed continued fraction has value $\frac{p}{q}$.

Utilizing signed continued fractions we define a minimum weight function μ: Rationals → Int by

$$\mu\left(\frac{p}{q}\right) = \min_{[a_0, a_1, \ldots, a_m] = \frac{p}{q}} \left\{ \sum_{i=0}^{m} \omega(a_i) \right\} , \qquad (8)$$

where ω is the minimum weight function over the integers [we shall later show $\mu\left(\frac{p}{1}\right) = \omega(p)$, so μ is an extention of ω from the integers to the rationals]. Letting $w[a_0, a_1, \ldots, a_m] = \sum_{i=0}^{m} \omega(a_i)$ denote the weight of the signed continued fraction $[a_0, a_1, \ldots, a_m]$, we further say $[a_0, a_1, \ldots, a_m] = \frac{p}{q}$

is a <u>minimum weight continued fraction</u> for $\frac{p}{q}$ whenever $w[a_0, a_1, \ldots, a_m] = \mu(\frac{p}{q})$.

Our goals for investigation of minimum weight continued fractions are

I. Find a formula and/or efficient procedure for determining $\mu(\frac{p}{q})$ for any rational p/q;

II. Provide a convenient and efficient algorithm for generating a minimum weight continued fraction $[a_0, a_1, \ldots, a_m]$ of value p/q for any p, q;

III. With $U(n) = \sum_{1 \le i, j \le n} \mu(\frac{i}{j})$, determine U(n) at least for small n, and find the asymptotic form of U(n).

Given the minimum weight continued fraction $[a_0, a_1, \ldots, a_m] = \frac{p}{q}$, the signed continued fraction $[a_1, a_2, \ldots, a_m] = \frac{q}{p - a_0 q}$ must also have minimum weight for $m \ge 1$. Thus using the fact that $\mu(p'/q') = \mu(q'/p')$ for $p' \neq 0$,

$$\mu(\frac{p}{q}) = \begin{cases} \omega(a_0) & \text{for } m = 0, \\ \omega(a_0) + \sum_{i=1}^{m} \omega(a_i) & \text{for } m \ge 1, \end{cases}$$

$$= \begin{cases} \omega(a_0) & \text{for } m = 0, \\ \omega(a_0) + \mu(\frac{q}{p - a_0 q}) & \text{for } m \ge 1, \end{cases}$$

$$= \omega(a_0) + \mu(\frac{p - a_0 q}{q}) \quad \text{for } m \ge 0.$$

Since a minimum weight continued fraction of value $\frac{p}{q}$ has $a_0 = i$ for some integer i, we obtain the identity

$$\mu(\tfrac{p}{q}) = \min_i \{ \omega(i) + \mu(\tfrac{p-iq}{q}) \} \quad \text{for any } \tfrac{p}{q}. \tag{9}$$

The identity (9) does not provide an effective algorithm for recursive computation of $\mu(\frac{p}{q})$, but the following sharper form of (9) does provide such an algorithm.

Theorem 7 For any $p > q \geq 1$, let $p = kq + r$ where $-q/2 < r \leq q/2$. Then

$$\mu(\tfrac{p}{q}) = \begin{cases} \omega(k) & \text{if } r = 0, \\ 1 + \min\{ \omega(k), \omega(k+1) \} & \text{if } r = q/2, \\ 1 + \min\{ \mu(\frac{r}{q}), \mu(\frac{q-r}{q}), \mu(\frac{p}{r}) \} & \text{if } 0 < r < q/2, \text{ and } k = 1, \\ \min\limits_{i=-1,0,1} \{ \omega(k+i) + \mu(\frac{iq-r}{q}) \} & \text{if } 0 < |r| < q/2, \text{ and } k \geq 2. \end{cases} \tag{10}$$

Assuming the validity of Theorem 7, note that the required values of μ for fractions p'/q' on the right hand side of (10) all have $q' \leq q$, $p' \leq p$, with $p' + q' < p + q$. Thus the recursion will not cycle and will allow the recursive computation of $\mu(p/q)$ from values of $\omega(j)$, j an integer, for any rational p/q.

The last line of the formula for $\mu(\frac{p}{q})$ dictates that we must consider three candidates, $\lfloor \frac{p}{q} + \frac{1}{2} \rfloor - 1$, $\lfloor \frac{p}{q} + \frac{1}{2} \rfloor$, and $\lfloor \frac{p}{q} + \frac{1}{2} \rfloor + 1$, as possible leading partial quotients to obtain a minimum weight continued fraction of value p/q. The fact that the two candidates $\lfloor \frac{p}{q} \rfloor$ and $\lceil \frac{p}{q} \rceil$ are not sufficient is confirmed by the following example. Note that $20/3 = [6, 1, 2] = [7, \overline{3}] = [8, \overline{1}, 4]$, where overbars denote negative partial quotients. Thus

$\mu(20/3) = w[8, \overline{1}, 4] = 3$, and since $\mu(2/3) = \mu(1/3) = 2$, the leading partial quotient of any minimum weight continued fraction for $20/3$ cannot be either $\lfloor 20/3 \rfloor = 6$ or $\lceil 20/3 \rceil = 7$.

In order to prove Theorem 7 we first note certain partial quotient transformation rules that preserve the value of the signed continued fraction. Letting the <u>length</u> of $[a_0, a_1, \ldots, a_m]$ denote the number of partial quotients, $m+1$, in the continued fraction, we shall in addition be concerned with how the transformation rules alter both the length and weight of the continued fraction.

For any real x, y, z with $z \neq 0$ and $y + \frac{1}{z} \neq 0$,

$$x + y + \frac{1}{z} = x + \cfrac{1}{0 + \cfrac{1}{y + \frac{1}{z}}}$$

and

$$x + z = x + \cfrac{1}{0 + \frac{1}{z}},$$

from which we obtain the following two rules.

<u>Rule 1. Internal Zero Deletion</u>

Given $[a_0, a_1, \ldots, a_{i-1}, 0, a_{i+1}, \ldots, a_m]$ for $1 \leq i \leq m-1$, then $[a_0, a_1, \ldots, a_{i-2}, (a_{i-1}+a_{i+1}), a_{i+2}, \ldots, a_m]$ is a signed continued fraction of the same value with no greater weight and length one less.

Rule 2. Partial Quotient Splitting

Given $[a_0, a_1, \ldots, a_m]$, then $[a_0, a_1, \ldots, a_{i-1}, a_i - a', 0, a', a_{i+1}, \ldots, a_m]$ is a signed continued fraction of the same value for any integer $a' \neq -\dfrac{1}{[a_{i+1}, \ldots, a_m]}$ for $i \leq m-1$, and any integer $a' \neq 0$ for $i = m$. This transformation increases the length by two, and the weight becomes greater than or equal to the previous weight.

Now let $u = \pm 1$, so $u^2 = 1$. Then for any real x, y, z with $z \neq 0$, and $y + \frac{1}{z} \neq 0$, and $u + (1/(y+1/z)) \neq 0$,

$$x + \cfrac{1}{u + \cfrac{1}{y + \frac{1}{z}}} = x + \frac{u(y + \frac{1}{z})}{y + u + \frac{1}{z}}$$

$$= x+u+ \cfrac{1}{-(y+u) + \cfrac{1}{-z}}$$

where we note $y + u + \frac{1}{z}$ cannot be zero. Furthermore, assuming $z \neq 0$ and $u + \frac{1}{z} \neq 0$,

$$x + \cfrac{1}{u + \frac{1}{z}} = x + u + \cfrac{1}{-(z+u)}$$

where also $z + u$ can not be zero. From these observations we obtain the result that any internal unit may be deleted from a signed continued fraction.

Rule 3. Internal Unit Deletion

Given $[a_0, a_1, \ldots, a_m]$ where $a_i = u = \pm 1$ for some particular i, $1 \leq i \leq m-1$, then $[a_0, a_1, \ldots, a_{i-2}, (a_{i-1}+u), (\overline{a_{i+1}+u}), \overline{a_{i+2}}, \ldots, \overline{a}_m]$ is a signed continued fraction of the same value with weight at most one greater and length one less.

Suppose $[a_0, a_1, \ldots, a_m]$ has $a_i = 2u$, where $u = \pm 1$, for some $1 \leq i \leq m-1$. Then by Rule 2 $[a_0, a_1, \ldots, a_{i-1}, u, 0, u, a_{i+1}, \ldots, a_m]$ has the same value if $[u, a_{i+1}, \ldots, a_m]$ is not zero. Furthermore, the units may be removed by two applications of Rule 3, so under the same assumptions $[a_0, a_1, \ldots, (a_{i-1}+u), \overline{2u}, (a_{i+1}+u), a_{i+2}, \ldots, a_m]$ also has the same value, from which we obtain Rule 4.

Rule 4. Internal Two Complementation

Given $[a_0, a_1, \ldots, a_m]$ where $a_i = 2u$ for $u = \pm 1$ for some particular i, $1 \leq i \leq m-1$, then $[a_0, a_1, \ldots, a_{i-2}, (a_{i-1}+u), \overline{2u}, (a_{i+1}+u), a_{i+2}, \ldots, a_m]$ is a signed continued fraction of the same value with the same length and with weight at most two greater whenever $[u, a_{i+1}, \ldots, a_m]$ is non zero, or equivalently whenever $[a_{i+1}, \ldots, a_m] \neq -u$.

For any p, q the signed continued fraction $[a_0, a_1, \ldots, a_m]$ of value either $\frac{p}{q}$ or $\frac{q}{p}$ is a minimal continued fraction whenever it has the minimum length over all minimum weight continued fractions of value either $\frac{p}{q}$ or $\frac{q}{p}$. Thus for any $\frac{p}{q}$, we have either $\frac{p}{q} = [a_0, a_1, \ldots, a_m]$ or $\frac{p}{q} = [0, a_0, a_1, \ldots, a_m]$ for some minimal $[a_0, a_1, \ldots, a_m]$.

The partial quotients of a minimal continued fraction satisfy several restrictive conditions which are enumerated in the following lemma. Let

$$E = \{k \mid k \neq 0,\ \omega(k+1) = \omega(k) + 1\},$$
$$F = \{k \mid k \neq 0,\ k \neq -2,\ \omega(k+1) \geq \omega(k)\}.$$

Lemma 8 For every minimal $[a_0, a_1, \ldots, a_m]$,

(i) $|a_i| \geq 1$ for $0 \leq i \leq m$, and $|a_m| \geq 2$ for $m \geq 1$;

(ii) $a_i = 1$ for $1 \leq i \leq m-1$ only if $a_{i-1}, a_{i+1} \in E$,
$a_i = -1$ for $1 \leq i \leq m-1$ only if $-a_{i-1}, -a_{i+1} \in E$;

(iii) $a_0 = 1$ only if $a_1 \in F$,
$a_0 = -1$ only if $-a_1 \in F$;

(iv) $|a_{i-1}| = |a_{i+1}| = 1$ for $1 \leq i \leq m-1$ only if
either $a_i/2 \in E$ and $a_{i-1} = a_{i+1} = 1$,
or $-a_i/2 \in E$ and $a_{i-1} = a_{i+1} = -1$;

(v) $a_i = 2,\ a_{i+1} = -2$ for $0 \leq i \leq m-1$ only if $a_{i-1}, -a_{i+2} \in F$,
$a_i = -2,\ a_{i+1} = 2$ for $0 \leq i \leq m-1$ only if $-a_{i-1}, a_{i+2} \in F$.

Proof Assume $[a_0, a_1, \ldots, a_m]$ is a minimal continued fraction of value value $\frac{p}{q}$. Now $|a_m| \neq 1$ for $m \geq 1$, since otherwise $[a_0, a_1, \ldots, a_{m-2}, (a_{m-1}+a_m)] = [a_0, \ldots, a_m]$ contradicts the minimality assumption. Now $a_i \neq 0$ for $1 \leq i \leq m-1$, since otherwise employing Rule 1 gives a contradiction, and $a_0 \neq 0$, since otherwise $[a_1, \ldots . a_m]$ has value $\frac{q}{p}$ and shorter length, a contradiction. This establishes assertion (i) of the lemma.

Assertion (ii) is immediate since otherwise application of Rule 3 deleting the unit would contradict the minimality of $[a_0, a_1, \ldots, a_m]$. When $a_0 = 1$, we may consider deletion of a_0 in $\frac{q}{p} = [0, a_0, a_1, \ldots, a_m]$ as well as deletion of any $a_i = \pm 1$ for $1 \leq i \leq m-1$ by Rule 3. Thus assertions (iii) and (iv) follow, for otherwise one or two applications of Rule 3 would yield a contradiction to the minimality assumption, e.g. $[1, \overline{2}, a_2, \ldots, a_m] = [0, 1, 1, \overline{a}_2, \ldots, \overline{a}_m] = [0, 2, a_2 - 1, a_3, \ldots, a_m] = 1/[2, a_2 - 1, a_3, \ldots, a_m]$ implies $a_0 = 1$, $a_1 = \overline{2}$ is not possible.

It follows from the minimality of $[a_0, a_1, \ldots, a_m]$ that $[a_{i+1}, a_{i+2}, \ldots, a_m]$ is minimal for any $0 \leq i \leq m-1$, so then $[a_{i+1}, a_{i+2}, \ldots, a_m] \neq \pm 1$ for any $0 \leq i \leq m-1$. Hence Rule 4 may be applied whenever $|a_i| = 2$ for $0 \leq i \leq m-1$. Assertion (v) then follows for all cases where $|a_{i-1}| \neq 2$, $|a_{i+2}| \neq 2$ since otherwise a single application of Rule 4 would contradict the minimality of $[a_0, a_1, \ldots, a_m]$. To see that the subsequences $2, \overline{2}, 2$ and $\overline{2}, 2, \overline{2}$ cannot occur in $[a_0, a_1, \ldots, a_m]$, one applies Rule 4 on the middle ± 2 followed by deletions of the neighboring resulting units by successive applications of Rule 3. This completes the proof of assertion (v) and the lemma.

□

Lemma 9 The value of any minimal $[a_0, a_1, \ldots, a_m]$ has the same sign as a_0 and the magnitude satisfies

$$|[a_0, a_1, \ldots, a_m]| > \begin{cases} \frac{2}{3} & \text{if } |a_0| = 1, \\ \frac{4}{3} & \text{if } |a_0| = 2, \\ \frac{7}{3} & \text{if } |a_0| \geq 3. \end{cases} \qquad (11)$$

Proof The assertion is immediate from Lemma 8 for all minimal $[a_0, a_1, \ldots, a_m]$ of length one or two. Proceeding by induction on the length of the minimal continued fractions, assume the assertion holds through length k for $k \geq 2$, and let $[a_0, a_1, \ldots, a_k]$ be minimal with length k+1. Without loss of generality we assume $a_0 \geq 1$, and note

$$[a_0, a_1, \ldots, a_k] = a_0 + \frac{1}{[a_1, a_2, \ldots, a_{k+1}]} \tag{12}$$

and

$$[a_0, a_1, \ldots, a_k] = a_0 + \frac{1}{a_1 + \frac{1}{[a_2, a_3, \ldots, a_{k+1}]}} \tag{13}$$

where $[a_1, \ldots, a_{k+1}]$ and $[a_2, \ldots, a_{k+1}]$ are minimal of length at most k, for which inequality (11) then holds by the induction assumption. Since $a_0 = |a_1| = 1$ is precluded by Lemma 8, from (11) and (12) we obtain $[a_0, a_1, \ldots, a_{k+1}] > 0$, so $[a_0, a_1, \ldots, a_{k+1}]$ has the same sign as a_0. Now if a_1 is also positive, inequality (11) is immediate so assume $a_1 \leq -1$, and from (11), (12) and (13),

$$[1, \overline{4}, a_2, \ldots, a_{k+1}] > 1 + \frac{1}{-4+\frac{3}{4}} = \frac{9}{13} > \frac{2}{3} \qquad \text{for} \begin{cases} a_2 \geq 2, \\ \text{or} \\ a_2 \leq -1, \end{cases}$$

$$[1, a_1, \ldots, a_{k+1}] > 1 + \frac{1}{-5 + \frac{3}{2}} = \frac{5}{7} > \frac{2}{3} \qquad \text{for} \quad a_1 \leq -5,$$

$$[2, \overline{2}, a_2, \ldots, a_{k+1}] > 2 + \frac{1}{-2 + \frac{3}{7}} = \frac{15}{11} > \frac{4}{3} \qquad \text{for} \begin{cases} a_2 \geq 4, \\ \text{or} \\ a_2 \leq -1, \end{cases}$$

$$[2, a_1, \ldots, a_{k+1}] > 2 - \frac{3}{7} = \frac{11}{7} > \frac{4}{3} \qquad \text{for} \quad a_1 \leq -3,$$

$$[3, \overline{2}, a_2, \ldots, a_{k+1}] > 3 + \frac{1}{-2 + \frac{3}{7}} = \frac{26}{11} > \frac{7}{3} \qquad \text{for} \quad a_2 \geq 3,$$

$$[3, a_1, \ldots, a_{k+1}] > 3 - \frac{3}{7} = \frac{17}{7} > \frac{7}{3} \qquad \text{for} \quad a_1 \leq -3,$$

$$[a_0, a_1, \ldots, a_{k+1}] > 4 - \frac{3}{2} > \frac{7}{3} \qquad \text{for} \quad a_0 \geq 4.$$

The preceeding seven cases are exhaustive by Lemma 8, verifying inequality (11) for $[a_0, a_1, \ldots, a_{k+1}]$. By induction the lemma is then proved.

□

Corollary 9.1 Let $[a_0, a_1, \ldots, a_m]$ be minimal of value z. Then a_0 has one of the at most three values $\lfloor z - \frac{1}{2} \rfloor$, $\lfloor z + \frac{1}{2} \rfloor$, $\lceil z + \frac{1}{2} \rceil$.

Proof For $m = 0$, the result is immediate. For $m \geq 1$, $z = a_0 + 1/[a_1, \ldots, a_m]$, so by Lemma 9, $|z - a_0| < \frac{3}{2}$, and the corollary follows.

□

Corollary 9.2 For every integer p,

$$\mu\left(\frac{p}{1}\right) = \omega(p), \tag{14}$$

$$\mu\left(2p \pm \frac{1}{2}\right) = 1 + \omega(p). \tag{15}$$

Proof The results (14) and (15) are immediate for $|p| \leq 1$. For $|p| \geq 2$, by Lemma 9 we must have $p = [a_0, a_1, \ldots, a_m]$ for some minimal $[a_0, \ldots, a_m]$. If $m = 0$ we obtain (14), so assume $m \geq 1$. By Corollary 9.1, $a_0 = p+1$ or $a_0 = p-1$, so $\omega(a_0) \geq \omega(p)-1$, and $\mu(\frac{p}{1}) = \sum_{i=0}^{m} \omega(a_i) \geq \omega(a_0) + \omega(a_1) \geq \omega(p)$, proving (14).

Since $[2p, \pm 2]$ has weight $1 + \omega(p)$, we obtain $\mu(2p \pm \frac{1}{2}) \leq 1 + \omega(p)$. For $|p| \geq 2$, both $2p - \frac{1}{2}$ and $2p + \frac{1}{2}$ are the values of some minimal $[a_0, a_1, \ldots, a_m]$ by Lemma 9 and by Corollary 9.1, a_0 is either $2p-1$, $2p$, or $2p+1$ in each case.

Since $\omega(2p) \leq \omega(2p \pm 1)$ holds for all p, we obtain $\mu(2k \pm \frac{1}{2}) \geq \omega(a_0) + \omega(a_1) \geq \omega(2p) + 1$, establishing (15).

□

We now proceed to complete the proof of Theorem 7.

Proof of Theorem 7

The first two cases of equation (10) are obtained from Corollary 9.2. For the fourth case let $p = kq + r$ with $0 < |r| < q/2$, where we assume $k \geq 2$. Then $\frac{p}{q} > \frac{3}{2}$, so by Lemma 9 $\frac{p}{q}$ is the value of some minimal $[a_0, a_1, \ldots, a_m]$. $|k - \frac{p}{q}| = \frac{|r|}{q} < \frac{1}{2}$ implies $k = \lceil \frac{p}{q} + \frac{1}{2} \rceil$, so by Corollary 9.1, a_0 must be either $k-1$, k, or $k+1$, verifying equation (10) for this case. The remaining case has $p = q + r$ with $0 < r < q/2$, hence $1 < \frac{p}{q} < \frac{3}{2}$. Using Lemma 9 we then have either a minimal $[a_0, a_1, \ldots, a_m]$ of value $\frac{p}{q}$ with $a_0 = 1$ or $a_0 = 2$, or value $\frac{q}{p}$ with $a_0 = 1$.

It follows that $\mu(\frac{p}{q}) = 1 + \min\{\mu(\frac{p}{q} - 1), \mu(\frac{p}{q} - 2), \mu(\frac{q}{p} - 1)\}$, which completes the proof of Theorem 7.

□

Now every $\frac{p}{q}$ has either a representation $\frac{p}{q} = [a_0, a_1, \ldots, a_m]$ or $\frac{p}{q} = [0, a_0, a_1, \ldots, a_m]$ where $[a_0, a_1, \ldots, a_m]$ is minimal. Then for any $0 \leq i \leq m$ using Lemma 9, $[a_i, a_{i+1}, \ldots, a_m] \neq 1/j$ for any integer j. These conditions assure that Rule 2 may be used recursively in a left to right order to generate a signed continued fraction $[b_0, b_1, \ldots, b_k]$ of value $\frac{p}{q}$, where $b_i \in \{0, \pm 1, \pm 2, \pm 2^2, \pm 2^3, \ldots\}$ for $0 \leq i \leq k$, and $[b_0, b_1, \ldots, b_k]$ has precisely $\mu(\frac{p}{q})$ non zero partial quotients.

In general, we say the signed continued fraction $[b_0, b_1, \ldots, b_m]$ is a <u>binary</u> continued fraction whenever $b_i = \pm 2^j$ or $b_i = 0$ for $0 \leq i \leq m$. Noting that the weight of a binary continued fraction is simply the number of its non zero terms, we have then proved the following assertion providing an alternative definition of $\mu(\frac{p}{q})$.

<u>Lemma 10</u> Every $\frac{p}{q}$ has a representation as a binary continued fraction, where $\mu(\frac{p}{q})$ is the minimum number of non zero partial quotients in any binary continued fraction of value $\frac{p}{q}$.

As an example, $[14, 29, 4]$ can be shown to be minimal, from which we obtain the minimum weight binary continued fraction $[16, 0, \overline{2}, 32, 0, \overline{4}, 0, 1, 4]$ of weight 6. Such a minimum weight representation is not unique even using the canonical signed digit form and ordering the partial quotients by decreasing powers of two for each initial b_i. Spe-

cifically, using Rules 2 and 3, $[14, 29, 4] = [14, 1, 0, 28, 4] = [15, \overline{1}, \overline{28}, \overline{4}]$ $= [16, 0, \overline{1}, \overline{1}, \overline{32}, 0, 4, \overline{4}]$, where the latter is also a minimum weight binary continued fraction.

A minimum weight representation of an integer may be efficiently determined recursively generating one "signed bit" per iteration. We now consider the possibility that a minimum weight binary continued fraction may be similarly determined.

For any rational z, b(z) is <u>admissible</u> for z as a leading binary partial quotient if $b(z) \in \{0, \pm 1, \pm 2, \pm 2^2, \pm 2^3, \ldots\}$, and $\mu(z) = 1 + \mu(z - b(z))$. Thus when b(z) is admissible for z and $[b_1, \ldots, b_m]$ is a minimum weight binary continued fraction of value $1/(z-b(z))$, we obtain $z = [b(z), b_1, b_2, \ldots, b_m]$ as a minimum weight representation of z. The sign and inversion symmetry of μ allow us to restrict our attention to the positive rationals greater than or equal to unity, where for $\frac{p}{q} \geq 4$ a well behaved solution is obtained.

<u>Lemma 11</u> For any rational number $z \geq 4$, $b(z) = 2^i$ is admissible for z whenever $\frac{23}{32} 2^i \leq z \leq \frac{25}{16} 2^i$.

<u>Proof</u> It follows from Lemma 8 that there is a minimal $[a_0, a_1, \ldots, a_m]$ of value z for any $z \geq 4$. First assume $8 \leq 2^i \leq z \leq \frac{25}{16} 2^i$. Then Then $z \leq \frac{25}{16} 2^i = \frac{5}{3} 2^i - \frac{5}{48} 2^i \leq \lceil \frac{5}{3} 2^i \rceil - \frac{3}{2}$, so $a_0 = z - \frac{1}{[a_1, a_2, \ldots, a_m]} \leq$ $\lceil \frac{5}{3} 2^i \rceil - \frac{3}{2} + \frac{3}{2}$, and then $2^i - 1 \leq a_0 \leq \lfloor \frac{5}{3} 2^i \rfloor$. Thus 2^i is a realizable leading signed bit for some minimum weight representation of the integer a_0, hence $\mu(z) = 1 + w[a_0 - 2^i, a_1, a_2, \ldots, a_m] = 1 + \mu(z - 2^i)$. Now consider

$4 \leq z \leq 6\ 1/4$, where then $z = [a_0, a_1, \ldots, a_m]$ for some minimal $[a_0, a_1, \ldots, a_m]$. Note that if $a_0 = 7$, then $|a_1| \neq 1$ and $z > 7 - \frac{2}{3}$, a contradiction. We obtain a similar contradiction assuming $a_0 = 3$. It follows that $4 \leq a_0 \leq 6$ and $b(z) = 4$ is then admissible for z. A similar argument for the regions $\frac{23}{32}\, 2^i \leq z \leq 2^i$ then completes the proof.

□

Hence the closest power of two is admissible for any $z \geq 4$. However, it follows from Theorem 7 that $\mu(\frac{22}{7}) = 3$ and that $[2, 1, \overline{8}]$ is the unique minimum weight binary continued fraction of value $\frac{22}{7}$, so 4 is not admissible for $\frac{22}{7}$. In fact, each element of the sequence $[2, 1, \overline{8}]$, $[2, 1, \overline{8}, 1, \overline{8}]$, $[2, 1, \overline{8}, 1, \overline{8}, 1, \overline{8}], \ldots$ with limiting value $[2, 1, \overline{8}, 1, \overline{8}, \ldots] = 6 - 2\sqrt{2}$ is a unique minimum weight binary continued fraction, and therefore represents a value where the closest power of two is not admissible. The next lemma shows that values in the interval $[\frac{3}{2}, 4]$ with the exceptional property that the closest power of two is not admissible are limited to relatively small neighborhoods around $\frac{3}{2}$ and 3.

Lemma 12 For the rational number z,

(i) $b(z) = 4$ is admissible whenever $4 \geq z \geq 6 - 2\sqrt{2} = 3.17157\ldots$,

(ii) $b(z) = 2$ is admissible whenever $2.82842\ldots = 2\sqrt{2} \geq z \geq 3 - \sqrt{2} = 1.58578\ldots$

Proof From Theorem 7 a minimal $[a_0, a_1, \ldots, a_m]$ of value z for $4 \geq z \geq 6 - 2\sqrt{2}$ must have $2 \leq a_0 \leq 5$, and $b(z) = 4$ is then admissible whenever $3 \leq a_0 \leq 5$. For $a_0 = 2$, note that since $[2, 1, \overline{8}, 1, \overline{8}, \ldots, 1, \overline{8}, 1, \overline{4}]$

$= [4, \overline{1}, \overline{4}, \overline{1}, \overline{4}, \ldots, \overline{1}, \overline{4}, \overline{1}, \overline{2}]$, it follows from Lemma 8 that any value less than four for which $b(z) = 4$ is not admissible must be less than $[2, 1, \overline{8}, 1, \overline{8}, 1, \overline{8}, \ldots] = 6 - 2\sqrt{2}$, proving (i). By a similar argument it follows for $\frac{3}{2} \leq z \leq 2\sqrt{2}$ that either $b(z) = 2$ or $b(z) = 1$ is admissible. Then since $[1, 2, \overline{4}, 2, \overline{4}, \ldots, 2, \overline{4}, 2, \overline{2}] = [\overline{2}, \overline{2}, \overline{2}, \overline{2}, \overline{2}, \ldots, \overline{2}, \overline{2}, \overline{2}, \overline{1}]$, it further follows from Lemma 8 that any value less than $2\sqrt{2}$ for which $b(z) = 2$ is not admissible must be less than $[1, 2, \overline{4}, 2, \overline{4}, \ldots] = 3 - \sqrt{2}$, proving (ii).

□

There is no result comparable to Lemmas 11 and 12 providing an interval range of z for which $b(z) = 1$ is admissible. To see why this is so note that $[2, \overline{2}, 4] = \frac{10}{7} = 1.4285 \ldots$, and $[2, \overline{2}, 4, \overline{2}, 4] = \frac{58}{41} = 1.4146 \ldots$, where it can be shown by Theorem 7 that 2 is the only admissible value for $\frac{10}{7}$ and $\frac{58}{41}$. Also $[0, 1, \overline{4}, 2, \overline{4}] = \frac{24}{17} = 1.4117 \ldots$, and $[0, 1, \overline{4}, 2, \overline{4}, 2, \overline{4}] = \frac{140}{99} = 1.4141 \ldots$, where 0 is the only admissible value for $\frac{24}{17}$ and $\frac{140}{99}$. Thus certain values approaching $\sqrt{2}$ from above have only 2 as admissible, and certain values approaching $\sqrt{2}$ from below have only 0 as admissible.

We are left with three regions of uncertainty relative to a direct minimum weight binary partial quotient generation process for $z \geq 1$, where in each of these regions it can be shown that at least one of two possible values must be admissible:

(i) for $2\sqrt{2} \leq z \leq 6 - 2\sqrt{2}$, either $b(z) = 2$ or $b(z) = 4$ is admissible,

(ii) for $\sqrt{2} \leq z \leq 3 - \sqrt{2}$, either $b(z) = 1$ or $b(z) = 2$ is admissible,

(iii) for $1 \leq z \leq \sqrt{2}$, either $b(z) = 0$ or $b(z) = 1$ is admissible.

We note that heuristics involving trial and some partial quotient look ahead for the above regions can be used to obtain "almost" minimum weight binary continued fractions that will bound from above and allow a close approximation of the average minimum weight binary continued fraction given as a function of n by $\left(\sum_{1 \leq p, q \leq n} \mu\left(\frac{p}{q}\right) \right) / n^2$.

It is also instructive to consider the minimum weight binary continued fraction representation of the successive convergents of particular irrational and transcendental numbers which are of essential importance to scientific computation. Now $\pi = [3, 7, 15, 1, 292, 1, 1, 1, 2, \ldots]$ with no known general pattern to the partial quotients. The leading partial quotients of a minimum weight binary representation of $[3, 7, 15, 1, 292, 1, 1, 1, 2]$ are then indicated in $[2, 1, \overline{8}, \overline{16}, 256, 0, 32, 0, 4]$ which can be used, if further truncated, to obtain a succession of very good approximations to π each computable by relatively few shift and add cycles as noted in the introduction.

A much stronger result is obtained for e which has the standard continued fraction representation $e = [2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, \ldots] = [2, (1, 2i, 1)_{i=1, 2, \ldots}]$.

Using the transformation rules we obtain $e = [2, 0, 1, (\overline{4}i, 2, 4i, 2)_{i=1, 2, \ldots}]$ which may be expanded to a binary continued fraction by inserting the minimum weight representation of 4i for each i. Note then that $[2, 0, 1, \overline{4}, 2, 4, 2, \overline{8}, 2] = \frac{2721}{1001} = 2.71828171$, which agrees with e to seven decimal places requiring only eight non zero binary partial quotients. Employing Theorem 6 it can be shown that, asymptotically, each additional non zero binary partial quotient digit in the above continued fraction reduces the error in the approximation of e by an average factor of 64, thus generating on the average "six bits of accuracy" per non zero bit of representation.

IV NUMERICAL RESULTS AND CONCLUSIONS

The recursion of Theorem 7 for computing the minimum weight function was utilized to obtain $\mu(\frac{p}{q})$ for all $1 \leq q < p \leq 2^{10}$. Values for $1 \leq q < p \leq 32$ are given in Table 2.

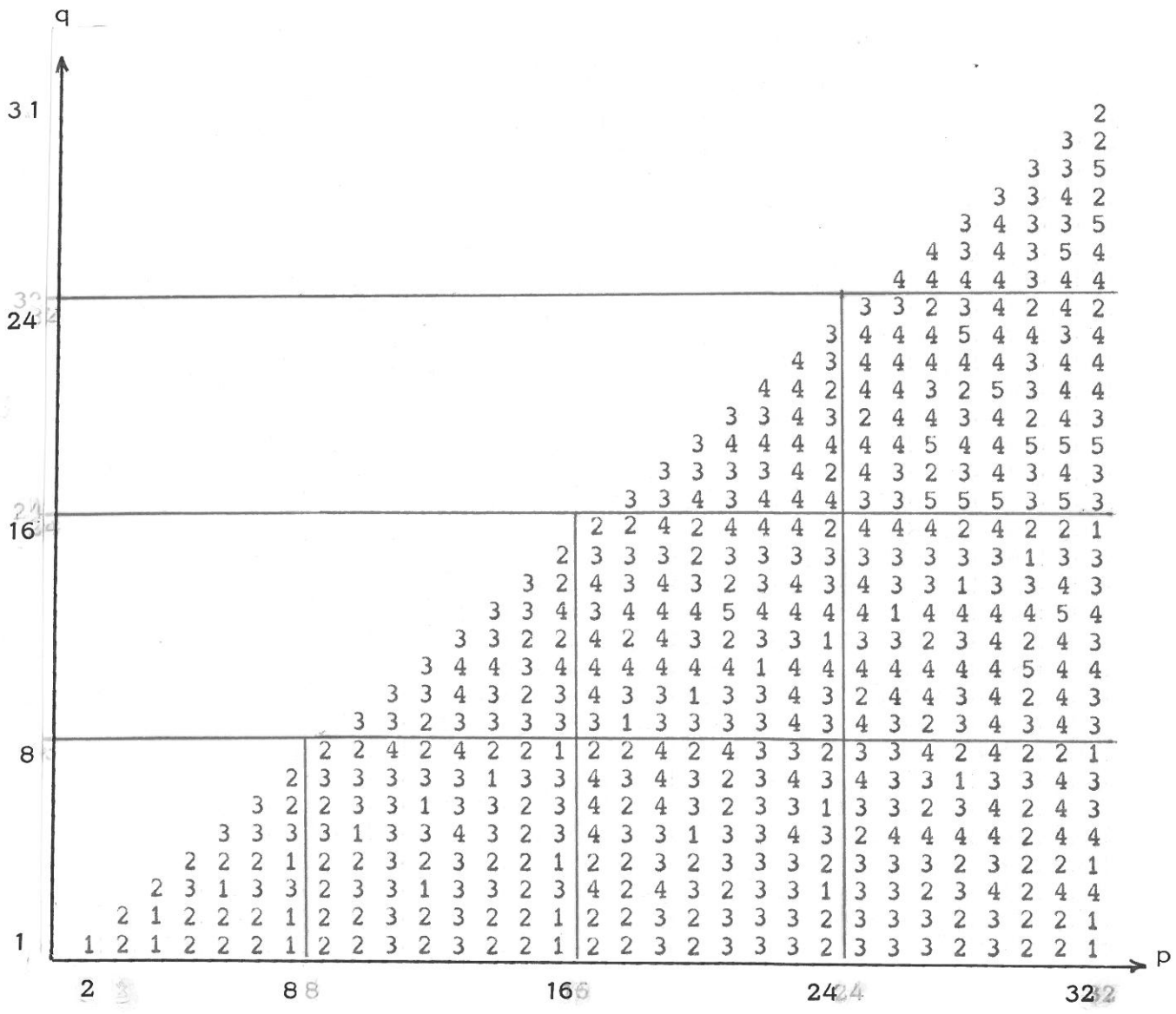


Table 2: Values of the minimum weight function $\mu(\frac{p}{q})$ for $1 \leq q < p \leq 32$.

Values of $U(n) = \sum_{1 \le p, q \le n} \mu(\frac{p}{q})$ and the average minimum weight $U(n)/n^2$ are tabulated in Table 3 for $n = 2^i$, $i = 1, 2, \ldots, 10$.

n	$U(n)$	$U(n)/n^2$	Δ	Δ
2^1	4	1.00000	-	-
2^2	22	1.37500	.37500	
2^3	120	1.87500	.50000	
2^4	602	2.35156	.47656	
2^5	3006	2.93555	.58399	
2^6	14358	3.50537	.56982	
2^7	67134	4.09753	.59216	
2^8	307880	4.69788	.60035	
2^9	1392148	5.31062	.61274	
2^{10}	6212770	5.92496	.61434	

Table 3: Values of $U(2^i) = \sum_{1 \le p \le q \le 2^i} \mu(\frac{p}{q})$, $U(2^i)/2^{2i}$, and $\Delta = U(2^i)/2^{2i} - U(2^{i-1})/2^{2i-2}$ for $i = 1, 2, \ldots, 10$.

From the established theory on the distribution of partial quotient values in standard simple continued fraction representation of real numbers, it is reasonable to expect $U(n)$ to have the following asymptotic form:

$$U(n) = cn^2 \log_2 n + o(n^2 \log_2 n). \qquad (16)$$

The value for Δ in Table 3 then must approach the constant c of (16), which would appear to give a value for c in the neighborhood of .62 . We have separately [MK81] determined the weight of the binary continued fraction representation of $\frac{p}{q}$ for each p, q in the range $1 \le p < q \le 2^{12}$

employing the "closest partial quotient" rule of choosing $a_0 = 2^k$ so as to minimize $|z - a_0|$ for all $z \geq 1$, even though this does not always give the the admissible minimum weight choice. Similar computation of values of Δ for $i = 1, 2, \ldots, 12$ show convergence, in this approximate minimum weight case, to $c^* = .660$, where further simulation runs for large samples with $1 \leq p < q \leq 2^i$ for i as large as 32 give results for c^* in the range .660 to .662 . Similarities in the second order differences (difference between successive Δ values) in Table 3 and in the "closest partial quotient" data strongly support the assumption that the value of c in (16) will be near .62 .

The limiting value of c is relevant to computer arithmetic for the following reasons:

(i) a value of $c < \frac{2}{3}$ implies that in the shift and add or subtract model of binary arithmetic, multiplication or division of fractions to a given level of average approximation error, using the minimum weight binary continued fraction representation for one argument, can be executed faster than floating point multiplication or division in the standard shift and add or subtract model, employing minimum weight integer representation for one argument, with results of comparable precision;

(ii) since the number of division cycles in the standard GCD algorithm for p, q (which equals the number of partial quotients of the standard simple continued fraction representation of $\frac{p}{q}$) converges to $0.5841 \log_2 n$ on the average for $1 \leq q < p \leq n$, a value of c in the

neighborhood of .62 in (16) implies that the shift and add or subtract model allows computation of the GCD by an average number of add or subtract operations only about 5% greater than the average number of division operations required in the standard Euclidian GCD algorithm.

Acknowledgment

We would like to thank Andy Freeman for preparation of programs to obtain the data of Tables 2 and 3.

References

[F61] Freiman, C.V., "Statistical analysis of certain binary division algorithms". Proc. IRE, vol. 49, pp. 91-103, 1961.

[G80] Gosper, unpublished manuscript.

[K69] Knuth, D.E., The Art of Computer Programming, Vol. 2, Reading, Mass.: Addison-Wesley, 1969.

[KM81] Kornerup, P. and Matula, D.W., in preparation.

[M61] MacSorley, O.L., "High speed arithmetic in binary computers", Proc. IRE, vol. 49, pp. 67-91, 1961.

[MK80] Matula, D.W., and Kornerup, P. "Foundations of Finite Precision Rational Arithmetic", Computing, Suppl. 2, pp. 85-111, 1980.

[MK81] Matula, D. W., and Kornerup, P., in preparation.

[M62] Metze, G., "A class of binary divisions yielding minimally represented quotients", IRE Trans. Electronic Computers, vol. EC-11, pp. 761-764, 1962.

[N56] Nadler, M., "A high speed electronic arithmetic unit for automatic computing machines", Acta Technica, no. 6, pp. 464-478, 1956.

[R60] Reitwiesner, G.W., "Binary arithmetic", in Advances in Computers, vol. 1, F.L. Alt, Ed. New York: Academic Press, 1960.

[R58] Robertson, J.E., "A new class of digital division methods", IRE Trans. Electronic Computers, vol. EC-7, pp. 218-222, 1958.

[R70] Robertson, J.E., "The Correspondence Between Methods of Digital Division and Multiplier Recoding Procedures", IEEE Trans. on Comp., vol. C-19, pp. 692-701, 1970.

[T58] Tocher, T.D., "Techniques of multiplication and division for automatic binary computers", Quart. J. Mech. Appl. Math., vol. 11, pt. 3, pp. 364-384, 1958.

[WL61] Wilson, J.B., and Ledley, "An Algorithm for Rapid Binary Division", IRE Trans., vol. EC-10, pp. 662-670, 1961.

PB-130

Matula & Kornerup: On Minimum Weight

TRYK: DAIMI/RECAU

ON MINIMUM WEIGHT BINARY REPRESENTATION OF INTEGERS AND CONTINUED FRACTIONS WITH APPLICATION TO COMPUTER ARITHMETIC *

by

David W. Matula
and
Peter Kornerup

DAIMI PB-130
January 1981

* This research was supported in part by the National Science Foundation under grant MCS-8012704.

Computer Science Department
AARHUS UNIVERSITY
Ny Munkegade – DK 8000 Aarhus C – DENMARK
Telephone: 06 – 12 83 55