48

ANALYSIS OF THE POPE-STEIN DIVISION ALGORITHM[1]

by

G. E. Collins[2]
and
D. R. Musser[3]

Technical Report #55

January 1969

[1] This research was supported in part by the National Science Foundation under Grant Number GJ-239, and by the U.S. Army Mathematics Research Center, under Contract No. DA-31-124-ARO-D-462.

[2] Also University of Wisconsin Computing Center.

[3] U. S. Army Mathematics Research Center Fellow.

1. INTRODUCTION

D. A. Pope and M. L. Stein in [4] described an algorithm for multiple-precision division in which trial quotient digits, relative to any base, are generated by dividing the two leading digits of the dividend by the one leading digit of the divisor. They showed that the trial quotient digits so obtained are never too small and that, if the divisor is normalized so that the leading digit is at least half the base, then the trial quotient digits are too large by at most two. In the Pope-Stein algorithm, the divisor is multiplied by the trial quotient digit, shifted, and subtracted from the dividend. If the result is non-negative, the trial quotient digit is correct; other wise it is too large so the divisor is shifted and added to the negative result. If this result is non-negative, the trial digit is one too large, while if it is still negative the trial digit is two too large and the divisor must be shifted and added a second time. Because an error in the trial digit entails so much additional computation, it is of interest to determine the respective probabilities, p_i of the possible errors i $(i = 0, 1, 2)$. These probabilities, which are functions of the base b and the number n of digits in the divisor, are computed in Section 3 of this paper. It is shown, in particular, that for $n \geq 2$ the limiting values, as b becomes large, are $p_0 = (77 - 6 \ln 2)/108 = .6745-$, $p_1 = (13 + 6 \ln 2)/54 = .3178-$ and $p_2 = (5 - 6 \ln 2)/108 = .0078-$.

Although these results reveal the Pope-Stein algorithm as being more efficient than one might have believed, we hasten to call attention to still more efficient division algorithms. Collins in [1] and [2] describes an algorithm for which the trial quotient digit is either correct or one too large, in which $p_1 \le 38/b$. This method uses a double-precision approximation to the inverse of the two leading digits of the divisor. Knuth in [3] describes another algorithm which is both simpler and more accurate. In the Knuth algorithm one starts with the Pope-Stein trial digit and makes a correction which is determined by the first three digits of the dividend and the first two digits of the divisor. The result is either correct or one too large and $p_1 \le 3/b$.

2. PROBLEM FORMULATION

Let $u = \Sigma_{i=0}^{m} u_i b^i$ and $v = \Sigma_{i=0}^{n-1} v_i b^i$ be positive integers in base b radix representation, where b is a large positive integer, $n \ge 2$ and $v_{n-1} > 0$. Actually we require only that $b > 6$, but since our main interest is in the case where b is (or is close to) the largest fixed point integer that can be stored in a single computer word, we will assume b is large and express most of our results in order notation.

We wish to compute $q = [u/v]$ and $r = u - qv$. If d is any positive integer, $\bar{u} = du$ and $\bar{v} = dv$, then $q = [\bar{u}/\bar{v}]$ and $r = (\bar{u} - q\bar{v})/d$. If we take $d = [b/(v_{n-1} + 1)]$ as in [3], then $[b/2]\, b^{n-1} \le \bar{v} < b^n$, i.e. $\bar{v}$ is normalized so that $\bar{v}_{n-1} \ge [b/2]$. For simplicity we henceforth assume b is even and we may therefore assume also that v has been normalized so that $v_{n-1} \ge b/2$.

We may assume $\Sigma_{i=m-n}^{m} u_i b^i < bv$, since otherwise we could set $u_{m+1} = 0$, replacing m by $m+1$. Then $q = \Sigma_{i=0}^{m-n} q_i b^i$ and it suffices to determine q_{m-n} since subsequent quotient digits will be determined in the same way (e.g. the process for q_{m-n-1} will be the same after u is replaced by $u - v \cdot q_{m-n} b^{m-n}$, where this new u will automatically satisfy $\Sigma_{i=m-n-1}^{m-1} u_i b^i < bv$), If $m > n$ then, as observed in [3], $q_{m-n} = [q/b^{m-n}] = [\,[u/v]/b^{m-n}] = [\,[u/b^{m-n}]/v]$, so it suffices to assume $m = n$ and determine $q = [u/v]$ where $q < b$.

In the Pope-Stein algorithm, the trial quotient, q', is computed by the formula

$$q' = \min(b-1,\ [(u_n b + u_{n-1})/v_{n-1}]\,).$$

p_i is then the probability that $q' - q = i$. These probabilities are not well-defined until we have defined a probability space. Our space E consists of the set of all triples (u, v, q) such that $b^n/2 \le v < b^n$, $0 \le q < b$ and $[u/v] = q$. Each point in E has probability $1/N$ where N is the number of points in E. Thus it is assumed that all possible normalized dividend-divisor combinations are equally likely.

3. PROBLEM SOLUTION

Letting E_i be the set of points in E such that $q' - q = i$, and N_i the number of points in E_i, we have $p_i = N_i/N$. It will suffice to determine N, N_0 and N_2, since $p_1 = 1 - p_0 - p_2$. We will assume $n \ge 2$ since trivially $p_0 = 1$, $p_1 = p_2 = 0$ when $n = 1$.

To compute N, the number of points in E, we note that the condition

$$0 \le [\tfrac{u}{v}] < b \tag{1}$$

is equivalent to $0 \le u < bv$, since u, v and b are integers. Thus for each value of v there are bv values of u such that (1) holds. Therefore

$$N = \sum_{v=b^n/2}^{b^n-1} bv = \frac{3}{8} b^{2n+1} - \frac{1}{4} b^{n+1} . \tag{2}$$

To compute N_0, the number of points in $E_0 = \{(u, v, q) \varepsilon E: q' = q\}$, we first assume $[u/v] = q \le b - 2$; then we must have

$$q' = [\frac{u_n b + u_{n-1}}{v_{n-1}}] = q \qquad (3)$$

in order for (u, v, q) to be in E_0. Equivalently,

$$q \le [u/b^{n-1}]/v_{n-1} < q + 1,$$

or

$$q v_{n-1} b^{n-1} \le u < (q+1) v_{n-1} b^{n-1} . \qquad (4)$$

From $[u/v] = q$ we also have $qv \le u < (q+1)v$; both this and (4) will be satisfied by u's which satisfy

$$qv \le u < (q+1)v_{n-1} b^{n-1} . \qquad (5)$$

Let $c = b^{n-1}$ and $v = jc + k$ (where $b/2 \le j < b$ and $0 \le k < c$). Then for a given v and q, the number of u's satisfying (5) is $\max(0, (q+1) v_{n-1} b^{n-1} - qv) = \max(0, (q+1) jc - q(jc+k)) = \max(0, jc - qk)$. Thus there are at least

$$N_{01} = \sum_{j=b/2}^{b-1} \sum_{k=0}^{c-1} \sum_{q=0}^{b-2} \max(0, jc - qk) \qquad (6)$$

points in E_0. But we must also consider the case where $q = b - 1$. Here, (3) is not required to hold; we know that $q^* \equiv [(u_n b + u_{n-1})/v_{n-1}] \ge b - 1$, and hence $q' = \min(q^*, b-1) = q$ even if $q^* \neq b - 1$. Thus we require only that $[u/v] = b - 1$, i.e. $v(b-1) \le u < vb$, and obtain $vb - v(b-1) = v$ points in E_0 for each v, for a total of

$$N_{02} = \sum_{v=b^n/2}^{b^n-1} v = \frac{3}{8} b^{2n} - \frac{1}{4} b^n . \qquad (7)$$

We now proceed to evaluate (6) in order to obtain $N_0 = N_{01} + N_{02}$.

In order to dispose of the max function in (6) we want to restrict the ranges of summation so that $jc - qk$ is always non-negative. To do this conveniently, and for later use, we introduce the sum N^*_{01} to be the same as N_{01} except that $b - 2$ replaces $b - 1$ as the upper limit on j . Also we reverse the order of the q and k summations. Thus

$$N_{01} = N^*_{01} + \sum_{q=0}^{b-2} \sum_{k=0}^{c-1} \{(b-1)b^{n-1} - qk\}, \qquad (8)$$

$$N^*_{01} = N_{011} + N_{012} ,$$

$$N_{011} = \sum_{j=b/2}^{b-2} \sum_{q=0}^{j} \sum_{k=0}^{c-1} (jc - qk) ,$$

$$N_{012} = \sum_{j=b/2}^{b-3} \sum_{q=j+1}^{b-2} \sum_{k=0}^{[jc/q]} (jc - qk) .$$

Evaluation of the sum in (8) and of N_{011} is a straightforward application of the formulas $\Sigma_{i=0}^{n} i = \frac{1}{2} n(n+1)$, $\Sigma_{i=0}^{n} i^2 = \frac{1}{6} n(n+1)(2n+1)$. We obtain

$$N_{01} = N^*_{01} + \frac{3}{4} b^{2n} + O(b^{2n-1}) \qquad (9)$$

$$N_{011} = \frac{7}{32} b^{2n+1} - \frac{3}{4} b^{2n} + O(b^{2n-1}) \qquad (10)$$

The formula for N_{011} in (10) holds only for $n > 2$; for $n = 2$ we have instead $N_{011} = \frac{7}{32} b^{2n+1} - \frac{65}{96} b^{2n} + O(b^{2n-1})$. To evaluate N_{012} we first perform the k summation (using a formula for the sum of an arithmetic series, $\Sigma_{i=0}^{n} (a + id) = (n+1)(a + \frac{1}{2} dn)$), and reverse the order of the q and j summations:

$$N_{012} = \sum_{q=b/2+1}^{b-2} \sum_{j=b/2}^{q-1} \left(\left[\frac{jc}{q}\right] + 1\right) \left(jc - \tfrac{1}{2} q \left[\frac{jc}{q}\right] \right)$$

Now, using the properties of the "floor" function $[x]$,

$$x - 1 < [x] \leq x < [x] + 1 ,$$

we let

$$S_1(q) = \sum_{j=b/2}^{q-1} \left(\frac{jc}{q}\right) \left(jc - \tfrac{1}{2} q \left(\frac{jc}{q}\right) \right)$$

$$S_2(q) = \sum_{j=b/2}^{q-1} \left(\frac{jc}{q} + 1\right) \left(jc - \tfrac{1}{2} q \left(\frac{jc}{q} - 1\right) \right)$$

so that

$$\sum_{q=b/2+1}^{b-2} S_1(q) < N_{012} < \sum_{q=b/2+1}^{b-2} S_2(q) \qquad (11)$$

We have

$$S_1(q) = \frac{c^2}{2q} \sum_{j=b/2}^{q-1} j^2$$

$$= \frac{c^2}{12q} \left[q - 1) q(2q - 1) - \left(\frac{b}{2} - 1\right) \frac{b}{2} (b - 1) \right]$$

$$= \frac{c^2}{12} \left[(q - 1)(2q - 1) - \frac{1}{4q} (b - 2) b(b - 1) \right] \qquad (12)$$

We first find

$$\sum_{q=b/2+1}^{b-2} (q - 1)(2q - 1) = \frac{7}{12} b^3 - \frac{35}{8} b^2 + O(b) . \qquad (13)$$

Then, letting $H_n = \Sigma_{i=1}^{n} \; i^{-1}$ and using the approximation

$$H_n = \ln n + \gamma + \frac{1}{2n} + O\left(\frac{1}{n^2}\right)$$

(where γ is Euler's constant), we obtain

$$\sum_{q=b/2+1}^{b-2} \frac{1}{q} = H_b - H_{b/2} - \frac{1}{b} - \frac{1}{b-1}$$

$$= \ln \frac{b}{b/2} + \frac{1}{2b} - \frac{1}{2(b/2)} - \frac{1}{b} - \frac{1}{b-1} + O(\frac{1}{b^2})$$

$$= \ln 2 - \frac{3}{2b} - \frac{1}{b-1} + O(\frac{1}{b^2}) \ . \qquad (14)$$

From (12-14) we have

$$\sum_{q=b/2+1}^{b-2} S_1(q) = \frac{c^2}{12}\{\frac{7}{12} b^3 - \frac{35}{8} b^2 + O(b) - \frac{1}{4}(b-2)b(b-1)(\ln 2 - \frac{3}{2b} - \frac{1}{b-1}$$

$$+ O(\frac{1}{b^2})) \}$$

$$= \frac{7-3\ln 2}{144} b^{2n+1} - \frac{5-\ln 2}{16} b^{2n} + O(b^{2n-1}) \ . \qquad (15)$$

Next we note that

$$S_2(q) = S_1(q) + \sum_{j=b/2}^{q-1} \{\frac{jc}{q}(\tfrac{1}{2}q) + jc - \tfrac{1}{2}q(\frac{jc}{q} - 1)\}$$

$$= S_1(q) + \sum_{j=b/2}^{q-1} (jc + \tfrac{1}{2}q)$$

$$= S_1(q) + c \cdot O(q^2) \ ,$$

and thus

$$\sum_{q=b/2+1}^{b-2} S_2(q) = \sum_{q=b/2+1}^{b-2} S_1(q) + O(b^{n+2}). \qquad (16)$$

We may now combine (11), (15) and (16) to conclude that N_{012} is equal to the right member of (15) when $n > 2$. When $n = 2$ we can only conclude that $N_{012} = \frac{7-3 \ln 2}{144} b^{2n+1} + O(b^{2n})$.

Finally, we can now compute, for $n > 2$,

$$N^*_{01} = N_{011} + N_{012} = \frac{77 - 6\ln 2}{288}\, b^{2n+1} - \frac{17 - \ln 2}{16}\, b^{2n} + O(b^{2n-1}) ,$$

$$N_{01} = N^*_{01} + \frac{3}{4}\, b^{2n} + O(b^{2n-1}) = \frac{77 - 6\ln 2}{288}\, b^{2n+1} - \frac{5 - \ln 2}{16}\, b^{2n} + O(b^{2n-1}) ,$$

and

$$N_0 = N_{01} + N_{02} = \frac{77 - 6\ln 2}{288}\, b^{2n+1} + \frac{1 + \ln 2}{16}\, b^{2n} + O(b^{2n-1}) .$$

Therefore, for $n > 2$,

$$p_0 = \frac{N_0}{N} = \frac{77 - 6\ln 2}{108} + \frac{1 + \ln 2}{6}\, b^{-1} + O(b^{-2}) ,$$

while

$$p_0 = \frac{77 - 6\ln 2}{108} + O(b^{-1}) \quad \text{for} \quad n = 2 .$$

The computation of N_2 is much easier, because N_2 can be shown to be related in a simple way to N^*_{01} . A point $(u, v, q) \varepsilon E$ is in E_2 if and only if $q \leq b - 3$ and $q^* - q = 2$. Given a $q \leq b - 3$ and a $v = jc + k$, then $(u, v, q) \varepsilon E_2$ for u satisfying

$$(q + 2)\, jc \leq u < (q + 1)\, v ,$$

and the number of such u's is $\max(0, (q+1)\,v - (q+2)jc) = \max(0, (q+1)k - jc)$.

Therefore

$$N_2 = \sum_{j=b/2}^{b-1} \sum_{k=0}^{c-1} \sum_{q=0}^{b-3} \max(0, (q+1)k - jc)$$

$$= \sum_{j=b/2}^{b-3} \sum_{q=j}^{b-3} \sum_{k=[\frac{jc}{q+1}]+1}^{c-1} ((q+1)k - jc)$$

$$= -\sum_{j=b/2}^{b-3} \sum_{q=j+1}^{b-2} \sum_{k=[jc/q]+1}^{c-1} (jc - qk)$$

$$= N_{012} - \sum_{j=b/2}^{b-3} \sum_{q=j+1}^{b-2} \sum_{k=0}^{c-1} (jc - qk)$$

$$= N_{012} + N_{011} - \sum_{j=b/2}^{b-2} \sum_{q=0}^{b-2} \sum_{k=0}^{c-1} (jc - qk)$$

$$= N_{01}^{*} - (\frac{1}{4} b^{2n+1} - b^{2n} + O(b^{2n-1})),$$

the evaluation of the last sum being straightforward. Thus, for $n > 2$,

$$N_2 = \frac{5 - 6 \ln 2}{288} b^{2n+1} - \frac{1 - \ln 2}{16} b^{2n} + O(b^{2n-1}),$$

and

$$p_2 = \frac{N_2}{N} = \frac{5 - 6 \ln 2}{108} - \frac{1 - \ln 2}{6} b^{-1} + O(b^{-2}).$$

For $n = 2$ we have $p_2 = \frac{5 - 6 \ln 2}{108} + O(b^{-1})$. Finally,

$$p_1 = 1 - p_0 - p_2 = \frac{13 + 6 \ln 2}{54} - \frac{\ln 2}{3} b^{-1} + O(b^{-2})$$

for $n > 2$ and $p_1 = \frac{13 + 6 \ln 2}{54} + O(b^{-1})$ for $n = 2$.

REFERENCES

[1] Collins, G. E. PM, A System for Polynomial Manipulation. CACM, 9 (1966), 578-589.

[2] Collins, G. E. The SAC-1 Integer Arithmetic System. Univ. of Wisconsin Computing Center Technical Reference (September 1967).

[3] Knuth, D. E. The Art of Computer Programming, Vol. 2 (Semi-Numerical Algorithms). Addison-Wesley, 1969.

[4] Pope, D. A., and Stein, M. L. Multiple Precision Arithmetic. CACM, 3 (1960), 652-654.